Article

11 TOPS photonic convolutional accelerator for optical neural networks

https://doi.org/10.1038/s41586-020-03063-0

Received: 15 April 2020

Accepted: 20 October 2020

Published online: 6 January 2021

Check for updates

Xingyuan Xu[1,9], Mengxi Tan[1], Bill Corcoran[2], Jiayang Wu[1], Andreas Boes[3], Thach G. Nguyen[3], Sai T. Chu[4], Brent E. Little[5], Damien G. Hicks[1,6], Roberto Morandotti[7,8], Arnan Mitchell[3] & David J. Moss[1]✉

Convolutional neural networks, inspired by biological visual cortex systems, are a powerful category of artificial neural networks that can extract the hierarchical features of raw data to provide greatly reduced parametric complexity and to enhance the accuracy of prediction. They are of great interest for machine learning tasks such as computer vision, speech recognition, playing board games and medical diagnosis[1–7]. Optical neural networks offer the promise of dramatically accelerating computing speed using the broad optical bandwidths available. Here we demonstrate a universal optical vector convolutional accelerator operating at more than ten TOPS (trillions (10^{12}) of operations per second, or tera-ops per second), generating convolutions of images with 250,000 pixels—sufficiently large for facial image recognition. We use the same hardware to sequentially form an optical convolutional neural network with ten output neurons, achieving successful recognition of handwritten digit images at 88 per cent accuracy. Our results are based on simultaneously interleaving temporal, wavelength and spatial dimensions enabled by an integrated microcomb source. This approach is scalable and trainable to much more complex networks for demanding applications such as autonomous vehicles and real-time video recognition.

Artificial neural networks are collections of nodes with weighted connections that, with proper feedback to adjust the network parameters, can 'learn' and perform complex operations for facial recognition, speech translation, playing strategy games and medical diagnosis[1–4]. Whereas classical fully connected feedforward networks face challenges in processing extremely high-dimensional data, convolutional neural networks (CNNs), inspired by the (biological) behaviour of the visual cortex system, can abstract the representations of input data in their raw form, and then predict their properties with both unprecedented accuracy and greatly reduced parametric complexity[5]. CNNs have been widely applied to computer vision, natural language processing and other areas[6,7].

The capability of neural networks is dictated by the computing power of the underlying neuromorphic hardware. Optical neural networks (ONNs)[8–12] are promising candidates for next-generation neuromorphic computation, because they have the potential to overcome some of the bandwidth bottlenecks of their electrical counterparts[6,13–15] such as for interconnections[16], and achieve ultrahigh computing speeds enabled by the >10-THz-wide optical telecommunications band[8]. Operating in analogue frameworks, ONNs avoid the limitations imposed by the energy and time consumed during reading and moving data back and forth for storage, known as the von Neumann bottleneck[13]. Important progress has been made in highly parallel, high-speed and trainable ONNs[8–12,17–22], including approaches that have the potential for full integration on a single photonic chip[8,12], in turn offering an ultrahigh computational density. However, there remain opportunities for substantial improvements in ONNs. Processing large-scale data, as needed for practical real-life computer vision tasks, remains challenging for ONNs because they are primarily fully connected structures and their input scale is determined solely by hardware parallelism. This leads to tradeoffs between the network scale and footprint. Moreover, ONNs have not achieved the extreme computing speeds that analogue photonics is capable of, given the very wide optical bandwidths that they can exploit.

Recently[22], the concept of time–wavelength multiplexing for ONNs was introduced and applied to a single perceptron operating at 11 billion (10^9) operations per second (giga-ops per second). Here, we demonstrate an optical convolutional accelerator (CA) to process and extract features from large-scale data, generating convolutions with multiple, simultaneous, parallel kernels. By interleaving wavelength, temporal and spatial dimensions using an integrated Kerr microcomb source[23–32], we achieve a vector computing speed as high as 11.322 TOPS. We then use it to process 250,000-pixel images, at a matrix processing speed of 3.8 TOPS.

The CA is scalable and dynamically reconfigurable. We use the same hardware to form both a CA front end and a fully connected neuron layer, and combine them to form an optical CNN. The CNN performs

[1]Optical Sciences Centre, Swinburne University of Technology, Hawthorn, Victoria, Australia. [2]Department of Electrical and Computer Systems Engineering, Monash University, Clayton, Victoria, Australia. [3]School of Engineering, RMIT University, Melbourne, Victoria, Australia. [4]Department of Physics, City University of Hong Kong, Tat Chee Avenue, Hong Kong, China. [5]Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. [6]Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. [7]INRS-Énergie, Matériaux et Télécommunications, Varennes, Québec, Canada. [8]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China. [9]Present address: Electro-Photonics Laboratory, Department of Electrical and Computer Systems Engineering, Monash University, Clayton, Victoria, Australia. ✉e-mail: dmoss@swin.edu.au

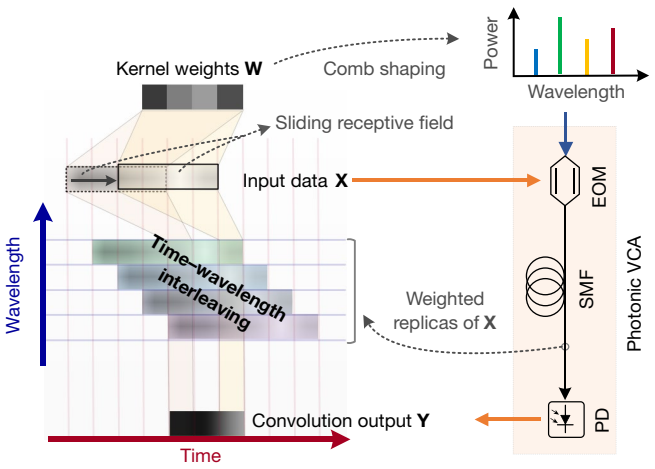


Fig. 1 | Operation principle of the TOPS photonic CA. EOM, electro-optical Mach–Zehnder modulator; SMF, standard single mode fibre for telecommunications; PD, photodetector.

simultaneous recognition of images from the MNIST handwritten digit dataset[33], achieving an accuracy of 88%. Our ONN represents a major step towards realizing monolithically integrated ONNs and is enabled by our use of an integrated microcomb chip. Moreover, the scheme is stand-alone and universal, fully compatible with either electrical or optical interfaces. Hence, it can serve as a universal ultrahigh-bandwidth front end that extracts data features for any neuromorphic hardware (optical or electronic-based), bringing massive-data machine learning for both real-time and ultrahigh-bandwidth data within reach.

Principle of operation

The photonic vector convolutional accelerator (VCA, Fig. 1) features high-speed electrical signal ports for data input and output. The input data vector **X** is encoded as the intensity of temporal symbols in a serial electrical waveform at a symbol rate $1/\tau$ (baud), where τ is the symbol period. The convolutional kernel is represented by a weight vector **W** of length R that is encoded in the optical power of the microcomb lines via spectral shaping by a waveshaper (see Methods). The temporal waveform **X** is then multi-cast onto the kernel wavelength channels via electro-optical modulation, generating the replicas weighted by **W**. The optical waveform is then transmitted through a dispersive delay with a delay step (between adjacent wavelengths) equal to the symbol duration of **X**, effectively achieving time and wavelength interleaving. Finally, the delayed and weighted replicas are summed via high-speed photodetection so that each time slot yields a convolution between **X** and **W** for a given convolution window, or receptive field.

As such, the convolution window effectively slides at the modulation speed matching the baud rate of **X**. Each output symbol is the result of R multiply-accumulate (MAC) operations, with the computing speed given by $2R/\tau$ TOPS. Since the speed of this process scales with both the baud rate and number of wavelengths, the massively parallel number of wavelengths from the microcomb yields speeds of many TOPS. Moreover, the length of the input data **X** is theoretically unlimited, so the CA can process data with an arbitrarily large scale—the only practical limitation being the external electronics.

Simultaneous convolution with multiple kernels is achieved by adding sub-bands of R wavelengths for each kernel. Following multicasting and dispersive delay, the sub-bands (kernels) are demultiplexed and detected separately, generating electronic waveforms for each kernel. The VCA is fully reconfigurable and scalable: the number and length of the kernels are arbitrary, limited only by the total number of wavelengths.

The CA processes vectors, which is extremely useful for human speech recognition or radio-frequency signal processing, for example. However, it can easily be applied to matrices for image processing by flattening the matrix into a vector. The precise way that this is performed is governed by the kernel size, which determines both the sliding convolution window's stride and the equivalent matrix computing speed. In our case the 3 × 3 kernel reduces the speed by a factor of 3, but we outline straightforward methods to avoid this (see Supplementary Information, including Supplementary Figs. 1–30 and Supplementary Tables 1, 2).

VCA and matrix CA

Extended Data Fig. 1 shows the experimental setup for the VCA, whereas Fig. 2 shows the matrix version used to process a classic 500 × 500 image, adopted from the University of Southern California-Signal and Image Processing Institute (USC-SIPI) database (http://sipi.usc.edu/database/). The system performs simultaneous image convolutions with ten 3 × 3 kernels. The weight matrices for all kernels were flattened into a composite kernel vector **W** containing all 90 weights (10 kernels with 3 × 3 = 9 weights each), which were then encoded onto the optical power of 90 microcomb lines by an optical spectral shaper (the waveshaper), with each kernel occupying its own band of 9 wavelengths. The wavelengths were supplied by a soliton crystal microcomb with a spacing of about 48.9 GHz (refs. [22–24,30,32]), with the 90 wavelengths occupying 36 nm across the C-band (see Extended Data Fig. 2 and Methods).

Figure 3 shows the image processing results. The 500 × 500 input image was flattened electronically into a vector **X** and encoded as the intensities of 250,000 temporal symbols, with a resolution of 8 bits per symbol (see Supplementary Information for a discussion on the effective number of bits), to form the electrical input waveform via a high-speed electrical digital-to-analogue converter, at a data rate of 62.9 gigabaud (with time slot τ = 15.9 ps; Fig. 3b). The duration of each image for all 10 kernels (3.975 μs) equates to a processing rate of 1/3.975 μs, or 0.25 million ultralarge-scale images per second.

The input waveform **X** was then multi-cast onto 90 shaped comb lines via electro-optical modulation, yielding replicas weighted by the kernel vector **W**. The waveform was then transmitted through around 2.2 km of standard single mode fibre (dispersion about 17 ps nm^{-1} km^{-1}) such that the relative temporal shift between the adjacent weighted wavelength replicas had a progressive delay of 15.9 ps, matching the data symbol duration τ. This resulted in time and wavelength interleaving for all 10 kernels. The 90 wavelengths were then de-multiplexed into 10 sub-bands of 9 wavelengths, with each sub-band corresponding to a kernel, and separately detected by 10 high-speed photodetectors. The detection process effectively summed the aligned symbols of the replicas (the electrical output waveform of kernel 4 is shown in Fig. 3c). The 10 electrical waveforms were converted into digital signals via analogue-to-digital converters and resampled so that each time slot of each individual waveform (wavelengths) corresponded to a dot product between one of the convolutional kernel matrices and the input image within a sliding window (that is, receptive field). This effectively achieved convolutions between the 10 kernels and the raw input image. The resulting waveforms thus yielded the 10 feature maps (convolutional matrix outputs) containing the extracted hierarchical features of the input image (Fig. 3d, Supplementary Information).

The VCA makes full use of time, wavelength and spatial multiplexing, where the convolution window effectively slides across the input vector **X** at a speed equal to the modulation baud rate of 62.9 billion symbols per second. Each output symbol is the result of 9 (the length of each kernel) MAC operations, and so the core vector computing speed (that is, the throughput) of each kernel is 2 × 9 × 62.9 = 1.13 TOPS. For 10 kernels the total computing speed of the VCA is therefore

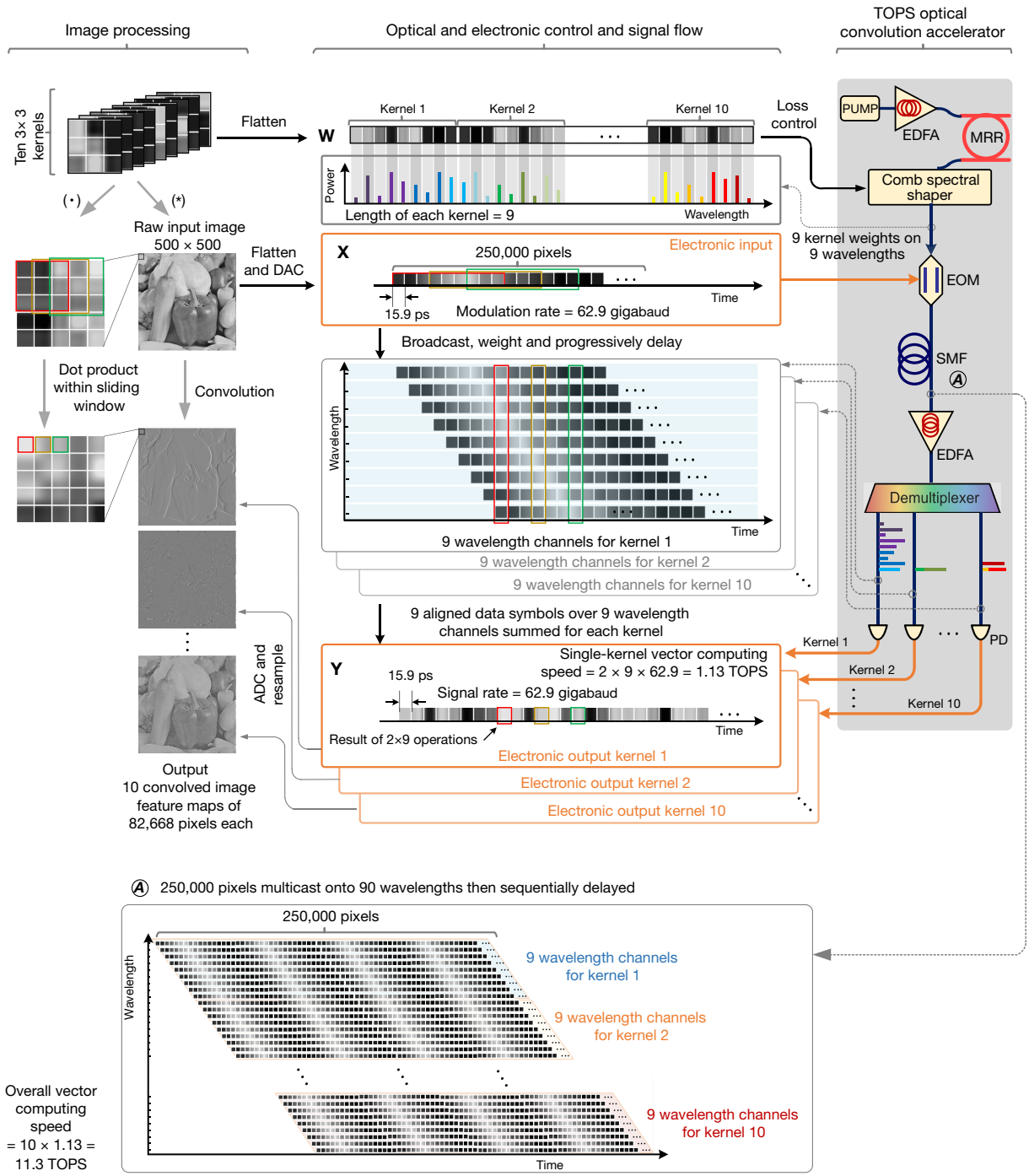


Fig. 2 | Image processing. The experimental setup (right panel), the optical and electronic control and signal flow (middle panel), and the corresponding processing flow of the raw input image (left panel) are shown. PUMP, continuous-wave pump laser; EDFA, erbium-doped fibre amplifier; MRR, micro-ring resonator. DAC, digital-to-analogue converter.

1.13 × 10 = 11.3 TOPS, representing a 500-fold increase for high-speed ONNs (see Supplementary Information).

In this work, the 3 × 3 kernels yielded a convolution window vertical sliding stride of 3, and so the effective matrix computing speed was 11.3/3 = 3.8 TOPS. Homogeneous strides operating at the full vector speed can be achieved by adding parallel weight-and-delay paths (see Supplementary Information), although we found that this was unnecessary. Finally, the CA can process arbitrarily large-scale data, beyond the 250,000 pixels reported here, limited only by the external electronics (compared to the single-neuron perceptron[22] that only processed 49-pixel images).

Convolutional ONN

Our CA is fully and dynamically reconfigurable as well as scalable. We used the same system to sequentially form both a front-end convolutional processor as well as a fully connected layer, to form an optical CNN that we applied to the simultaneous recognition of full 10 (0 to 9) handwritten digit images[33], versus two-digit recognition[22].

Extended Data Figure 3 shows the principle of the optical CNN, whereas Fig. 4 shows the experimental configuration. The convolutional layer performs the heaviest computing duty of the network, taking 55% to 90% of the total computing power. The digit images

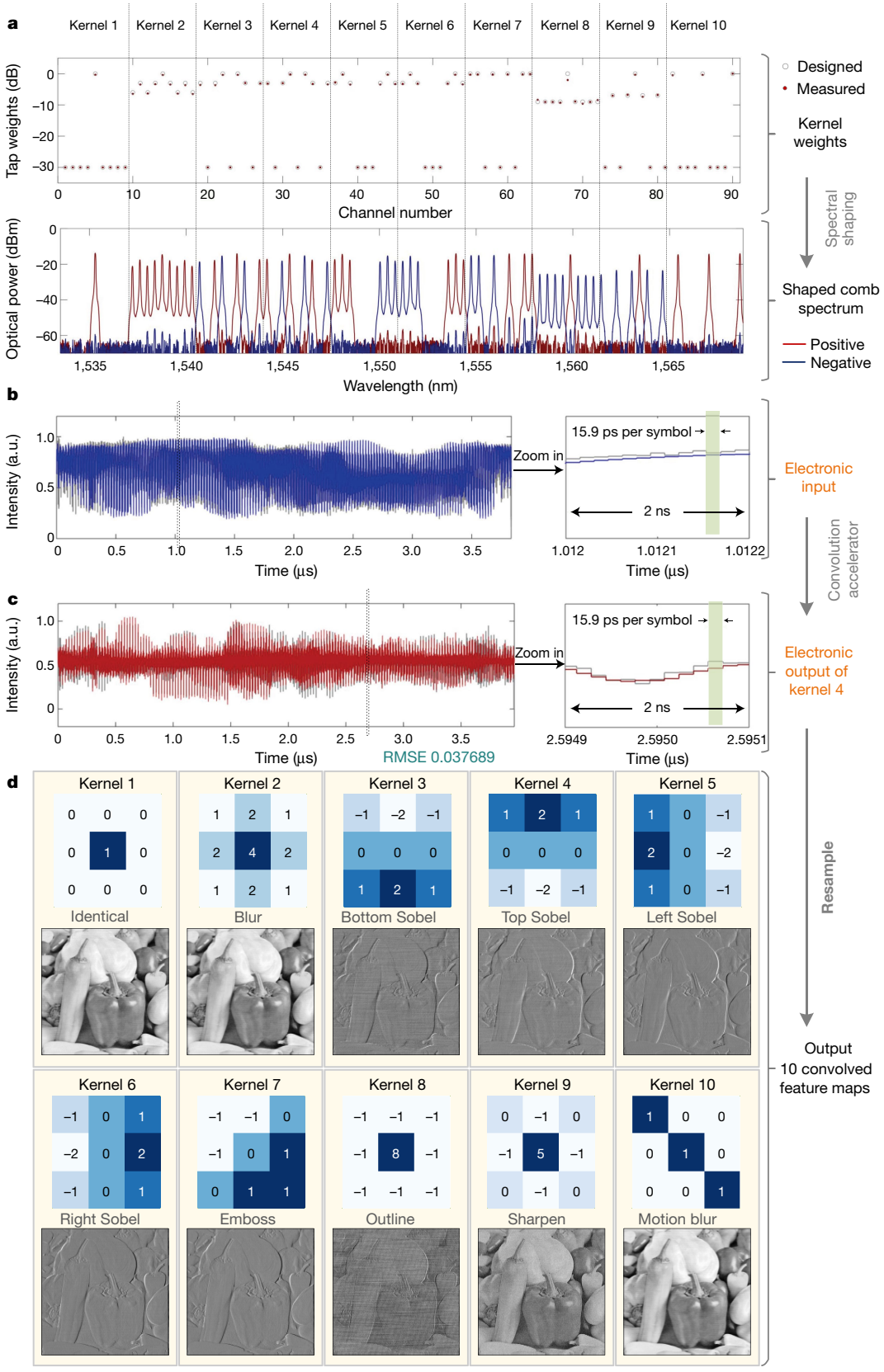


Fig. 3 | Experimental results of the image processing. a, The kernel weights (tap weights) and the shaped microcomb's optical spectrum. **b**, The input electrical waveform of the image (the grey and blue lines show the ideal and experimentally generated waveforms, respectively). **c**, The convolved results of the fourth kernel that performs a top Sobel image processing function—a gradient-based method that looks for strong changes in the first derivative of an image (the grey and red lines show the ideal and experimentally generated waveforms, respectively). **d**, The weight matrices of the kernels and corresponding recovered images. RMSE, root-mean-square error.

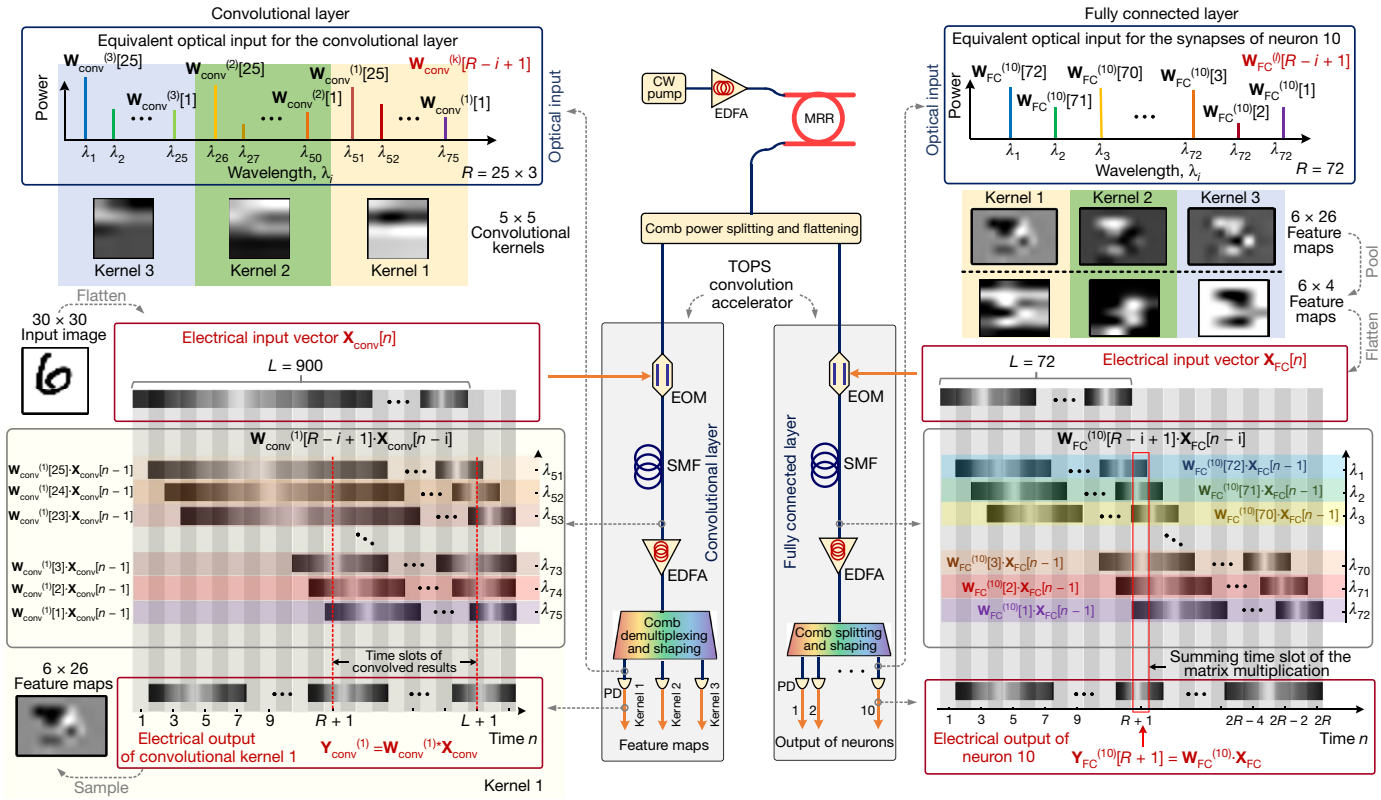


Fig. 4 | Experimental schematic of the optical CNN. Left side is the input front-end CA while the right side is the fully connected layer, both of which form the deep learning optical CNN. The microcomb source supplies the wavelengths for both the TOPS photonic CA as well as the fully connected layer systems. The electronic digital signal processing (DSP) module used for sampling and pooling and so on is external to this structure.

(30 × 30 greyscale matrices with 8-bit resolution) were flattened into vectors and multiplexed in the time domain at 11.9 gigabaud (time slot $\tau = 84$ ps). Three 5 × 5 kernels were used, requiring 75 microcomb lines (Fig. 5) and resulting in a vertical convolution stride of 5. The dispersive delay was achieved with around 13 km of standard single mode fibre for telecommunications to match the data baud rate. The wavelengths were de-multiplexed into the three kernels which were detected by high-speed photodetectors and then sampled as well as nonlinearly scaled with digital electronics. This recovered the hierarchical feature maps that were then pooled electronically and flattened into a vector $\mathbf{X}_{FC}$ (72 × 1) for each image, forming the input data to the fully connected layer.

The fully connected layer had 10 neurons, one for each of the 10 categories of handwritten digits (0 to 9), with the synaptic weights of the lth neuron ($l \in [1, 10]$) represented by a 72 × 1 weight matrix $\mathbf{W}_{FC}^{(l)}$. The number of comb lines (72) matched the length of the flattened feature map vector $\mathbf{X}_{FC}$. The shaped optical spectrum at the lth port had an optical power distribution proportional to the weight vector $\mathbf{W}_{FC}^{(l)}$, serving as the optical input for the lth neuron. After being multicast onto 72 wavelengths and progressively delayed, the optical signal was weighted and demultiplexed with a single waveshaper into 10 spatial output ports, each corresponding to a neuron. Since this part of the network involved linear processing, the kernel wavelength weighting could be implemented either before electro-optical modulation or later—that is, just before photodetection. The advantage of the latter is that both demultiplexing and weighting can be achieved with a single waveshaper. Finally, the different node/neuron outputs were obtained by sampling the 73rd symbol of the convolved results. The final output of the optical CNN was represented by the intensities of the output neurons (Extended Data Fig. 4), where the highest intensity for each tested image corresponded to the predicted category. The peripheral systems, including signal sampling, nonlinear function and pooling, were implemented electronically with digital signal processing hardware, although many of these functions (for example, pooling) can be achieved optically (for example, with the VCA). Supervised network training was performed offline electronically (see Supplementary Information).

We first experimentally tested 50 images of the handwritten digit MNIST dataset[33], followed by more extensive testing on 500 images (see Supplementary Information for 500 image results). The confusion matrix for 50 images (Fig. 6) shows an accuracy of 88% for the generated predictions, in contrast to 90% for the numerical results calculated on an electrical digital computer. The corresponding results for 500 images are essentially the same—89.6% for theory versus 87.6% for experiment (Supplementary Fig. 25). The fact that the CNN achieved close to the theoretical accuracy indicates that the impact of effects that could limit the network performance and reduce the effective number of bits (see Supplementary Information), such as electrical and optical noise or optical distortion (owing to high-order dispersion), is small.

The computing speed of the VCA front end of the optical CNN was 2 × 75 × 11.9 = 1.785 TOPS. For processing the image matrices with 5 × 5 kernels, the convolutional layer had a matrix flattening overhead of 5, yielding an image computing speed of 1.785/5 = 357 billion operations per second. The computing speed of the fully connected layer was 119.8 billion operations per second (see Supplementary Information). The waveform duration was 30 × 30 × 84 ps = 75.6 ns for each image, and so the convolutional layer processed images at the rate of 1/75.6 ns = 13.2 million handwritten digit images per second. The optical CNN supports online training, given that the dynamic reconfiguration response time of the optical spectral shaper used to establish the synapses is <500 ms, and even faster with integrated optical spectral shapers[34].

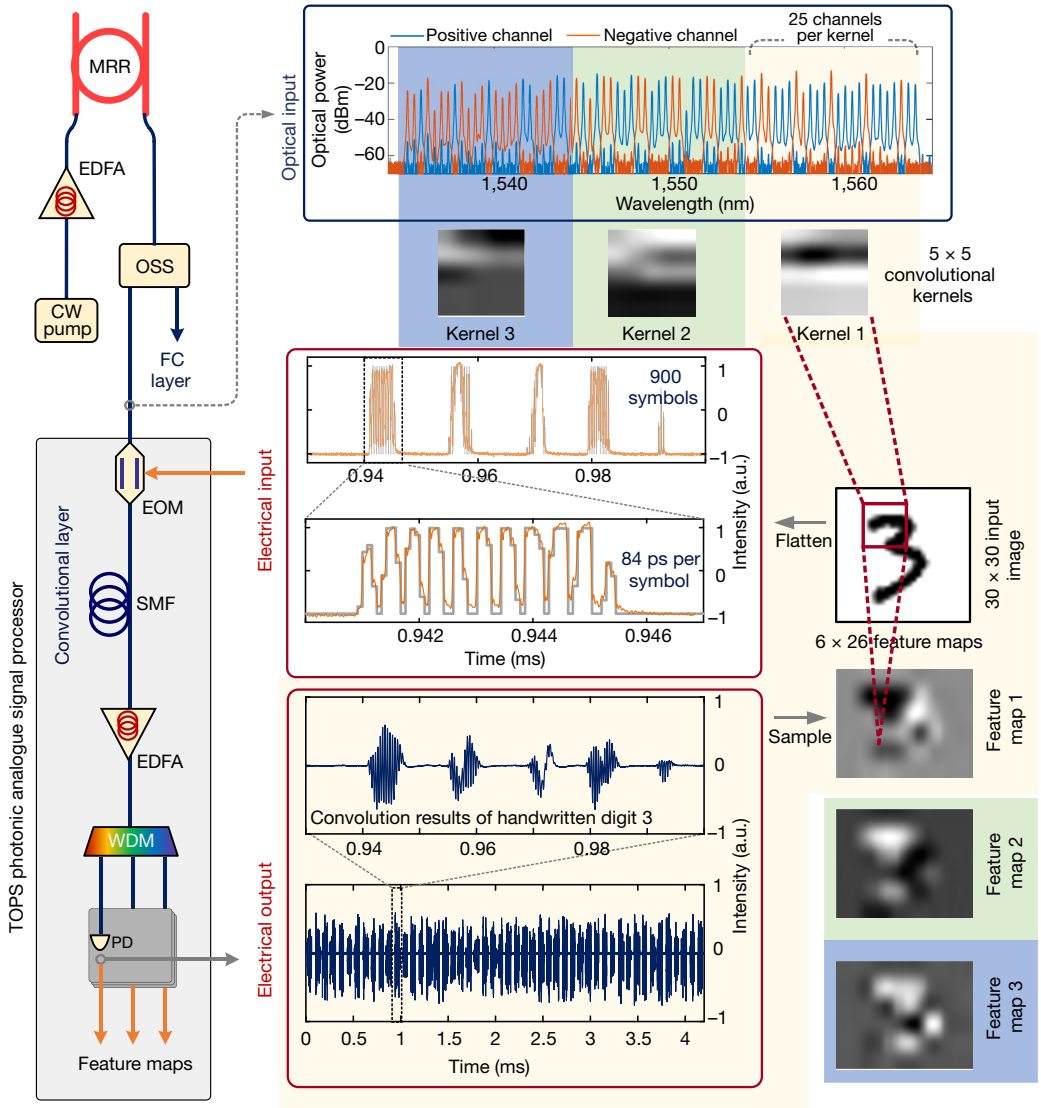


Fig. 5 | Convolutional layer. Architecture and experimental results. The left panel shows the experimental setup. The right panel shows the experimental results of one of the convolutional kernels, showing the shaped microcomb's optical spectrum and the corresponding kernel weights (the blue and red lines denote the positive and negative synaptic weights, respectively), the input electrical waveform for the digit 3 (middle: the grey and yellow lines show the ideal and experimentally generated waveforms, respectively), the convolved results and the corresponding feature maps. CW pump, continuous-wave pump laser. OSS, optical spectral shaper. WDM, wavelength de-multiplexer. FC, fully connected.

Although handwritten digit recognition is a common benchmark for digital hardware, it is still largely beyond current analogue reconfigurable ONNs. Digit recognition requires many physical parallel paths for fully connected networks (for example, a hidden layer with 10 neurons requires 9,000 physical paths), which represents a huge challenge for nanofabrication. Our CNN represents the first reconfigurable and integrable ONN capable not only of performing high-level complex tasks such as full handwritten digit recognition, but to do so at many TOPS.

Discussion

Although the performance of ONNs is not yet competitive with leading-edge electronic processors at >200 TOPS (for example, Google TPU[15] and other chips[13,14,35]), there are straightforward approaches towards increasing our performance both in scale and speed (see Supplementary Information). Further, with a single processor speed of 11.3 TOPS, our VCA is approaching this range. The CA is fundamentally limited in data size only by the electrical digital-to-analogue converter memory, and processing 4K-resolution (4,096 × 2,160 pixels) images at >7,000 frames per second is possible.

The 720 synapses of the CNN (72 wavelengths per synapses per neuron, 10 neurons), a substantial increase for optical networks[12], enabled us to classify the MNIST dataset[33]. Nonetheless, further scaling is needed to increase the theoretical prediction accuracy from 90% to that of state-of-the-art electronics, typically substantially greater than 95% (see Supplementary Information). Both the CA and CNN can be scaled substantially in size and speed using only off-the-shelf telecommunications components. The full S, C and L telecommunications bands (1,460–1,620 nm, >20 THz) would allow more than 400 channels (at a 50-GHz spacing), with the further ability to use polarization and spatial dimensions, ultimately enabling speeds beyond a quadrillion (10^{15}) operations per second (peta-ops per second) with more than 24,000 synapses for the CNN (see Supplementary Information). Supplementary Fig. 30 shows theoretical results for a scaled network that achieves an accuracy of 94.3%. This can, in principle, be further increased to achieve accuracies comparable to state-of-the-art electronic chips for the tasks performed here.

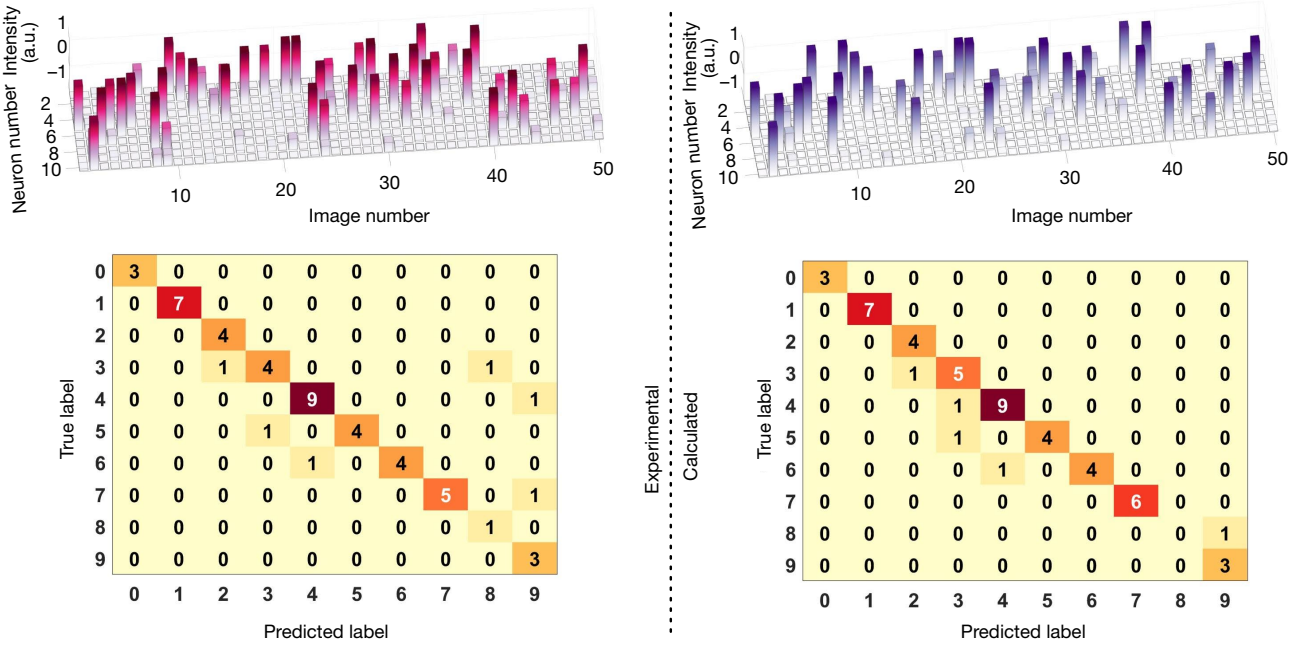


Fig. 6 | Experimental and theoretically calculated results for image recognition. The upper figures show the sampled intensities of the 10 output neurons at the fully connected layer, while the lower figures show the confusion matrices (see Supplementary Information), with the darker colours indicating a higher recognition score.

Although our systems had a non-negligible optical latency of 0.11 μs introduced by the propagation delay of the dispersive fibre spool, this did not affect the operation speed. Moreover, this latency is not fundamental: it can almost be eliminated (to <200 ps) by using integrated highly dispersive devices such as photonic crystals or customized chirped Bragg gratings[36] that can achieve the required differential progressive delay time (15.9 ps). Further, propagation loss as low as 0.7 dB m^{-1} (ref. [37]) has now been reported in Si_3N_4 waveguides, allowing lengths >10 m (corresponding to 50 ns), sufficient for low-loss integrated gratings.

Finally, current nanofabrication techniques can enable much higher levels of integration of the CA. The microcomb source itself is based on a CMOS (complementary metal–oxide–semiconductor)-compatible platform, amenable to large-scale integration. Other components such as the optical spectral shaper, modulator, dispersive media, de-multiplexer and photodetector have all been realized in integrated (albeit simpler) forms[36–38].

Optical neuromorphic processing is a comparatively young field, yet ONNs have now reached the TOPS regime, with the potential to reach the regime of peta-ops per second. Therefore, we are optimistic that optical neuromorphic hardware will eventually play an important part in computationally intensive operations, perhaps in a support role within a hybrid opto-electronic framework. This will help to alleviate the heavy computing cost of digital electronics, while enhancing the overall performance of neuromorphic processors.

Conclusion

We have demonstrated a universal optical convolutional accelerator operating at 11.3 TOPS for vector processing, and use a matrix-based approach to perform convolutions of large-scale images with 250,000 pixels. We then use the same hardware to form an optical CNN for the recognition of the full 10-digit set of handwritten images. Our network is capable of processing large-scale data and images at ultrahigh computing speeds for real-time massive-data machine learning tasks, such as identifying faces in cameras as well as assessing pathology for clinical scanning applications[39,40].

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/s41586-020-03063-0.

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Schalkoff, R. J. Pattern recognition. In *Wiley Encyclopedia of Computer Science and Engineering* (ed. Wah, B. W.) https://doi.org/10.1002/9780470050118.ecse302 (Wiley, 2007).
3. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
4. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
5. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
6. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
7. Lawrence, S., Giles, C. L., Tsoi, A. C. & Back, A. D. Face recognition: a convolutional neural-network approach. *IEEE Trans. Neural Netw.* **8**, 98–113 (1997).
8. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
9. 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).
10. Peng, H.-T., Nahmias, M. A., de Lima, T. F., Tait, A. N. & Shastri, B. J. Neuromorphic photonic integrated circuits. *IEEE J. Sel. Top. Quantum Electron.* **24**, 6101715 (2018).
11. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
12. 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).
13. Ambrogio, S. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60–67 (2018).
14. Esser, S. K. et al. Convolutional networks for fast, energy-efficient neuromorphic computing. *Proc. Natl Acad. Sci. USA* **113**, 11441–11446 (2016).
15. Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471–476 (2016).
16. Miller, D. A. B. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
17. Appeltant, L. et al. Information processing using a single dynamical node as complex system. *Nat. Commun.* **2**, 468 (2011).
18. 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).

19. Vandoorne, K. et al. Experimental demonstration of reservoir computing on a silicon photonics chip. *Nat. Commun*. **5**, 3541 (2014).
20. Brunner, D., Soriano, M. C., Mirasso, C. R. & Fischer, I. Parallel photonic information processing at gigabyte per second data rates using transient states. *Nat. Commun*. **4**, 1364 (2013).
21. Tait, A. N., Chang, J., Shastri, B. J., Nahmias, M. A. & Prucnal, P. R. Demonstration of WDM weighted addition for principal component analysis. *Opt. Express* **23**, 12758–12765 (2015).
22. Xu, X. et al. Photonic perceptron based on a Kerr microcomb for high-speed, scalable, optical neural networks. *Laser Photon. Rev*. **14**, https://doi.org/10.1002/lpor.202000070 (2020).
23. Pasquazi, A. et al. Micro-combs: a novel generation of optical sources. *Phys. Rep*. **729**, 1–81 (2018).
24. Moss, D. J., Morandotti, R., Gaeta, A. L. & Lipson, M. New CMOS-compatible platforms based on silicon nitride and Hydex for nonlinear optics. *Nat. Photon*. **7**, 597–607 (2013).
25. Kippenberg, T. J., Gaeta, A. L., Lipson, M. & Gorodetsky, M. L. Dissipative Kerr solitons in optical microresonators. *Science* **361**, eaan8083 (2018).
26. Savchenkov, A. A. et al. Tunable optical frequency comb with a crystalline whispering gallery mode resonator. *Phys. Rev. Lett*. **101**, 093902 (2008).
27. Spencer, D. T. et al. An optical-frequency synthesizer using integrated photonics. *Nature* **557**, 81–85 (2018).
28. Marin-Palomo, P. et al. Microresonator-based solitons for massively parallel coherent optical communications. *Nature* **546**, 274–279 (2017).
29. Kues, M. et al. Quantum optical microcombs. *Nat. Photon*. **13**, 170–179 (2019).
30. Cole, D. C., Lamb, E. S., Del'Haye, P., Diddams, S. A. & Papp, S. B. Soliton crystals in Kerr resonators. *Nat. Photon*. **11**, 671–676 (2017).
31. Stern, B., Ji, X., Okawachi, Y., Gaeta, A. L. & Lipson, M. Battery-operated integrated frequency comb generator. *Nature* **562**, 401–405 (2018).
32. Wu, J. et al. RF photonics: an optical microcombs' perspective. *IEEE J. Sel. Top. Quant. Electron*. **24**, 6101020 (2018).
33. LeCun, Y., Cortes, C. & Borges, C. J. C. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/
34. Metcalf, A. J. et al. Integrated line-by-line optical pulse shaper for high-fidelity and rapidly reconfigurable RF-filtering. *Opt. Express* **24**, 23925–23940 (2016).
35. NVIDIA Corporation. *Comparison of Convolution Methods for GPUs*. http://ska-sdp.org/publications/released-sdp-memos-2 (2018).
36. Sahin, E., Ooi, K., Png, C. & Tan, D. Large, scalable dispersion engineering using cladding-modulated Bragg gratings on a silicon chip. *Appl. Phys. Lett*. **110**, 161113 (2017).
37. Roeloffzen, C. G. H. et al. Low-loss Si_3N_4 TriPleX optical waveguides: technology and applications overview. *IEEE J. Sel. Top. Quantum Electron*. **24**, 4400321 (2018).
38. Wang, C. et al. Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages. *Nature* **562**, 101–104 (2018).
39. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
40. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).

Methods

Optical soliton crystal microcomb

Optical frequency combs, composed of discrete and equally spaced frequency lines, are extremely powerful tools for optical frequency metrology[23]. Microcombs offer the full power of optical frequency combs, but in an integrated form with much smaller footprint[23–25]. They have enabled many breakthroughs through their ability to generate wideband low-noise optical frequency lines for high-resolution optical frequency synthesis[27], ultrahigh-capacity communications[28], complex quantum state generation[29], advanced microwave signal processing[32], and more.

In this work we use a particular class of microcomb termed soliton crystals. They were so named because of their crystal-like profile in the angular domain of tightly packed self-localized pulses within micro-ring resonators[30]. They are naturally formed in micro-cavities with appropriate mode crossings, without the need for complex dynamic pumping and stabilization schemes (described by the Lugiato–Lefever equation[23]). They are characterized by distinctive 'fingerprint' optical spectra (Extended Data Fig. 2f) which arise from spectral interference between the tightly packaged solitons circulating along the ring cavity. This category of soliton microcomb features deterministic soliton formation originating from the mode-crossing-induced background wave and the high intracavity power (the mode crossing is measured as in Extended Data Fig. 2c). This in turn enables simple and reliable initiation via adiabatic pump wavelength sweeping[29] that can be achieved with manual detuning (the intracavity power during the pump sweeping is shown in Extended Data Fig. 2d). The ability to adiabatically sweep the pump lies in the fact that the intra-cavity power is over thirty times higher than for single-soliton states (dissipative Kerr solitons), and very close to that of spatiotemporal chaotic states[23]. Thus, the soliton crystal displays much less thermal detuning or instability resulting from the 'soliton step' that makes resonant pumping of dissipative Kerr soliton states more challenging[23]. It is this combination of ease of generation and overall conversion efficiency that makes soliton crystals highly suited for demanding applications such as ONNs.

The coherent soliton crystal microcomb (Extended Data Fig. 2) was generated by optical parametric oscillation in a single integrated MRR. The MRR (Extended Data Fig. 2b) was fabricated on a CMOS-compatible doped silica platform[23,24], featuring a high Q factor of over 1.5 million and a radius of 592 μm, which corresponds to a low free spectral range of about 48.9 GHz (ref. [32]). The pump laser (Yenista Tunics, 100S-HP) was boosted by an optical amplifier (Pritel PMFA-37) to initiate the parametric oscillation. The soliton crystal microcomb provided more than 90 channels over around 36 nm of the telecommunications C-band (1,540–1,570 nm), offering adiabatically generated low-noise frequency comb lines with a small footprint of <1 mm^2 and potentially low power consumption (<100 mW using the technique in ref. [31]).

Evaluation of the computing performance

Given that there are no common standards in the literature for classifying and quantifying the computing speed and processing power of ONNs, we explicitly outline the performance definitions that we use in characterizing our performance. We follow an approach that is widely used to evaluate electronic micro-processors. The computing power of the CA—closely related to the operation bandwidth—is denoted as the throughput, which is the number of operations performed within a certain period. Considering that in our system the input data and weight vectors originate from different paths and are interleaved in different dimensions (time, wavelength and space), we use the temporal sequence at the electrical output port to define the throughput in a more straightforward manner.

At the electrical output port, the output waveform has $L+R-1$ symbols in total (L and R are the lengths of the input data vector and the kernel weight vector, respectively), of which $L-R+1$ symbols are the convolution results. Further, each output symbol is the calculated outcome of R MAC operations or $2R$ operations, with a symbol duration τ given by that of the input waveform symbols. Thus, considering that L is generally much larger than R in practical CNNs, the term $(L-R+1)/(L+R-1)$ would not affect the vector computing speed, or throughput, which (in operations per second) is given by

$$\frac{2R}{\tau}\cdot\frac{L-R+1}{L+R-1}\approx\frac{2R}{\tau} \quad (1)$$

As such, the computing speed of the VCA demonstrated here is $2\times 9\times 62.9\times 10 = 11.321$ TOPS for 10 parallel convolutional kernels).

We note that when processing data in the form of vectors, such as audio speech, the effective computing speed of the VCA would be the same as the vector computing speed $2R/\tau$. Yet when processing data in the form of matrices, such as for images, we must account for the overhead on the effective computing speed brought about by the matrix-to-vector flattening process. The overhead is directly related to the width of the convolutional kernels, for example, with 3×3 kernels, the effective computing speed would be approximately $1/3\times 2R/\tau$, which, however, we note is still in the ultrafast (TOPS) regime owing to the high parallelism brought about by the time–wavelength interleaving technique.

For the matrix CA the output waveform of each kernel (with a length of $L-R+1 = 250{,}000-9+1 = 249{,}992$) contains $166\times 498 = 82{,}668$ useful symbols that are sampled out to form the feature map, while the rest of the symbols are discarded. As such, the effective matrix convolution speed for the experimentally performed task is slower than the vector computing speed of the CA by the overhead factor of 3, and so the net speed then becomes $11.321\times 82{,}668/249{,}991 = 11.321\times 33.07\% = 3.7437$ TOPS.

For the deep CNN the CA front-end layer has a vector computing speed of $2\times 25\times 11.9\times 3 = 1.785$ TOPS while the matrix convolution speed for 5×5 kernels is $1.785\times 6\times 26/(900-25+1) = 317.9$ billion operations per second. For the fully connected layer of the deep CNN (see Supplementary Information), the output waveform of each neuron would have a length of $2R-1$, while the useful (relevant output) symbol would be the one located at $R+1$, which is also the result of $2R$ operations. As such, the computing speed of the fully connected layer would be $2R/(\tau\times(2R-1))$ per neuron. With $R = 72$ during the experiment and 10 neurons operating simultaneously, the effective computing speed of the matrix multiplication would be $2R/(\tau\times(2R-1))\times 10 = 2\times 72/(84\text{ ps}\times(2\times 72-1)) = 119.83$ billion operations per second.

Experiment

To achieve the designed kernel weights, the generated microcomb was shaped in power using liquid-crystal-on-silicon-based spectral shapers (Finisar WaveShaper 4000S). We used two waveshapers in the experiments—the first was used to flatten the microcomb spectrum while the precise comb power shaping required to imprint the kernel weights was performed by the second, located just before the photodetection. A feedback loop was employed to improve the accuracy of comb shaping, in which the error signal was generated by first measuring the impulse response of the system with a Gaussian pulse input and comparing it with the ideal channel weights. (The shaped impulse responses for the convolutional layer and the fully connected layer of the CNN are shown in the Supplementary Information).

The electrical input data were temporally encoded by an arbitrary waveform generator (Keysight M8195A, 65 billion symbols per second, 25-GHz analogue bandwidth) and then multicast onto the wavelength channels via a 40-GHz intensity modulator (iXblue). For the 500×500 image processing, we used sample points at a rate of 62.9 billion samples per second to form the input symbols. We then employed a 2.2-km-long dispersive fibre that provided a progressive delay of 15.9 ps per channel, precisely matched to the input baud rate. For the convolutional layer of the CNN, we used 5 sample points at 59.421642 billion samples

per second to form each single symbol of the input waveform, which also matched with the progressive time delay (84 ps) of the 13-km-long dispersive fibre (the generated electronic waveforms for 50 images are shown are shown in the Supplementary Information; these served as the electrical input signal for the convolutional and fully connected layers, respectively). We note that the high-order dispersion present in standard single mode fibre would introduce an alignment error into the convolution results (up to 46 ps). Thus, we used the programmable phase characteristics of the second waveshaper to compensate for this error. This could equally be addressed by using speciality dispersive elements with negligible high-order dispersion.

For the CA in both experiments—the 500 × 500 image processing experiment and the convolutional layer of the CNN—the second waveshaper simultaneously shaped and de-multiplexed the wavelength channels into separate spatial ports according to the configuration of the convolutional kernels. As for the fully connected layer, the second waveshaper simultaneously performed the shaping and power splitting (instead of de-multiplexing) for the 10 output neurons. Here, we note that the de-multiplexed or power-split spatial ports were sequentially detected and measured. However, these two functions could readily be achieved in parallel with a commercially available 20-port optical spectral shaper (WaveShaper 16000S, Finisar) and multiple photodetectors.

The negative channel weights were achieved using two methods. For the 500 × 500 image processing experiment and the convolutional layer of the CNN, the wavelength channels of each kernel were separated into two spatial outputs by the waveshaper, according to the signs of the kernel weights, and then detected by a balanced photodetector (Finisar XPDV2020). Conversely, for the fully connected layer the weights were encoded in the symbols of the input electrical waveform during the electrical digital processing stage. Incidentally, we demonstrate the possibility of using different methods to impart negative weights, both of which work in the experiments.

Finally, the electrical output waveform was sampled and digitized by a high-speed oscilloscope (Keysight DSOZ504A, 80 billion symbols per second) to extract the final convolved output. In the CNN, the extracted outputs of the CA were further processed digitally, including rescaling to exclude the loss of the photonic link via a reference symbol, and then mapped onto a certain range using a nonlinear tanh function. The pooling layer's functions were also implemented digitally, following the algorithm introduced in the network model.

The residual discrepancy or inaccuracy in our work for both the recognition and convolving functions, as compared to the numerical calculations, was due to the deterioration of the input waveform caused by intrinsic limitations in the performance of the electrical arbitrary waveform generator. Addressing this would readily lead to a higher degree of accuracy (that is, closer agreement with the numerical calculations).

During the experimental testing of 500 handwritten digit images using the CNN, we separated the input dataset into 10 batches (each batch contains 50 images), which were sequentially processed by the convolutional layer and fully connected layer (see Supplementary Information for the full results).

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files.

Code availability

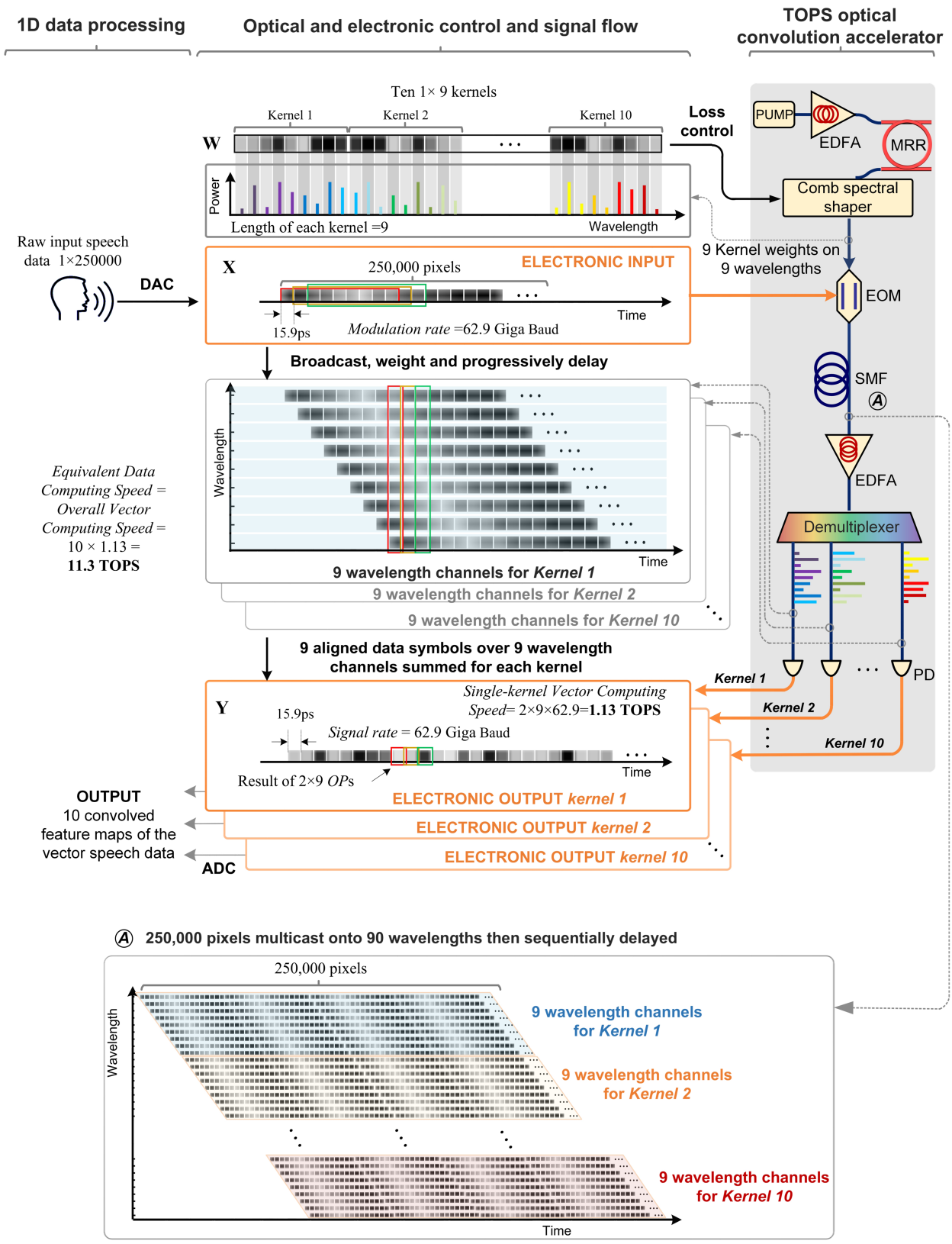
The authors declare that the algorithm of the demonstrated neural network supporting the findings of this study is available within the paper and its supplementary information files.

Acknowledgements This work was supported by the Australian Research Council Discovery Projects Program (grant numbers DP150104327, DP190102773 and DP190101576). R.M. acknowledges support by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Strategic, Discovery and Acceleration Grants Schemes, by the MESI PSR-SIIRI Initiative in Quebec, and by the Canada Research Chair Program. B.E.L. was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number XDB24030000). D.G.H. was supported in part by the Australian Research Council (grant number FT104101104). R.M. is affiliated with the Institute of Fundamental and Frontier Sciences (China) as an adjoint faculty member.

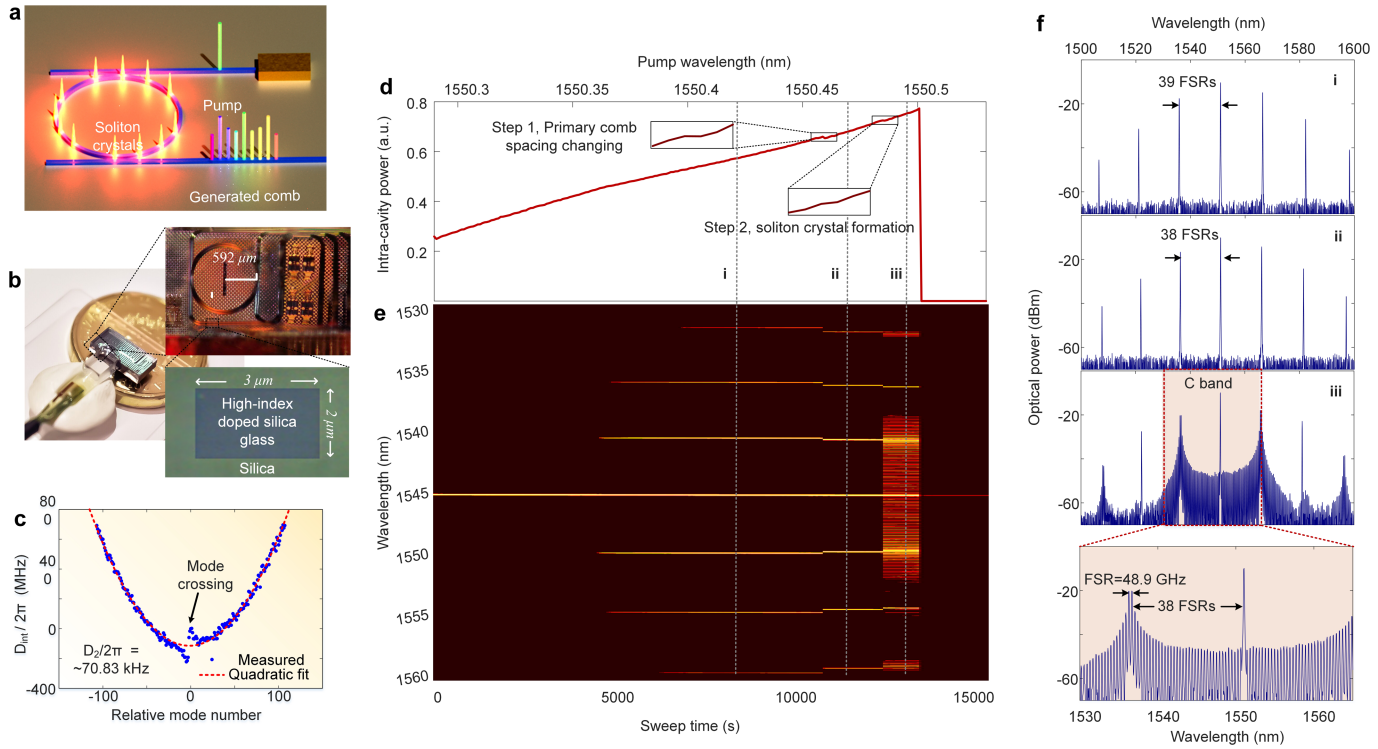
Author contributions X.X. conceived the idea and designed the project. X.X. and M.T. performed the experiments. X.X. analysed the data, and performed the numerical simulations and the offline training. S.T.C. and B.E.L. designed and fabricated the integrated devices. B.C., J.W., A.B., T.G.N., R.M. and A.M. contributed to the development of the experiment and to the data analysis. X.X. and D.J.M. wrote the manuscript. D.J.M. supervised the research.

Competing interests The authors declare no competing interests.

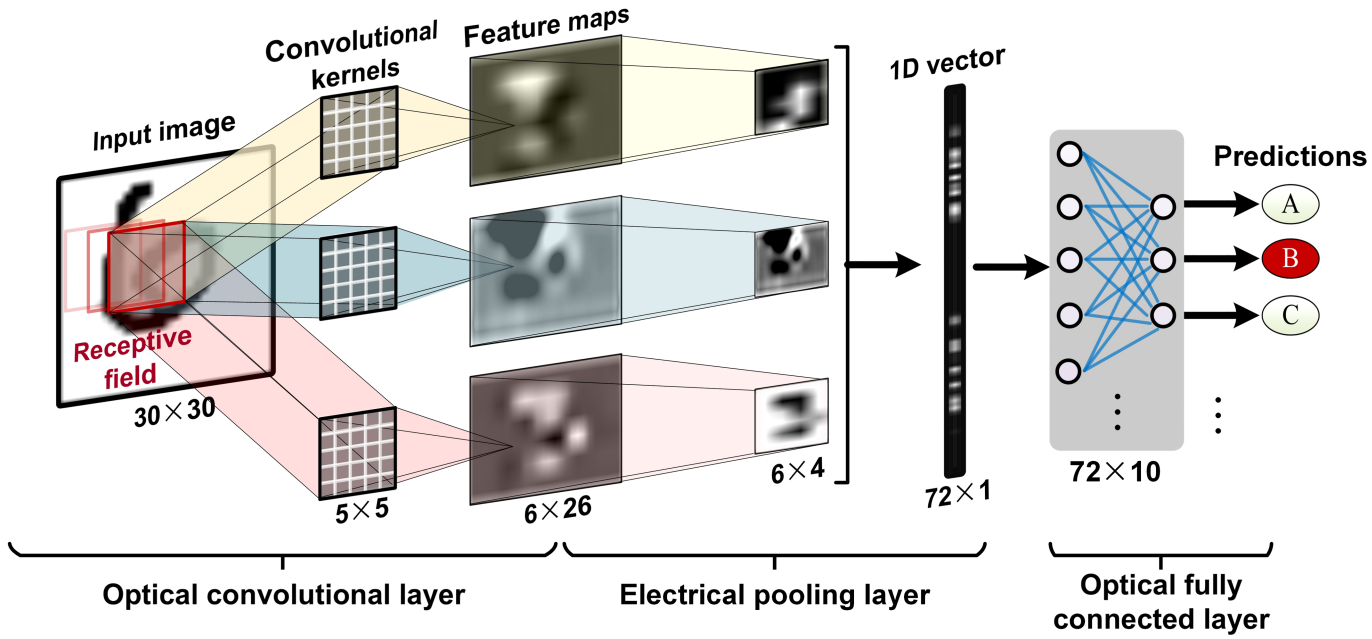
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-03063-0.
Correspondence and requests for materials should be addressed to D.J.M.
Peer review information *Nature* thanks Sylvain Gigan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



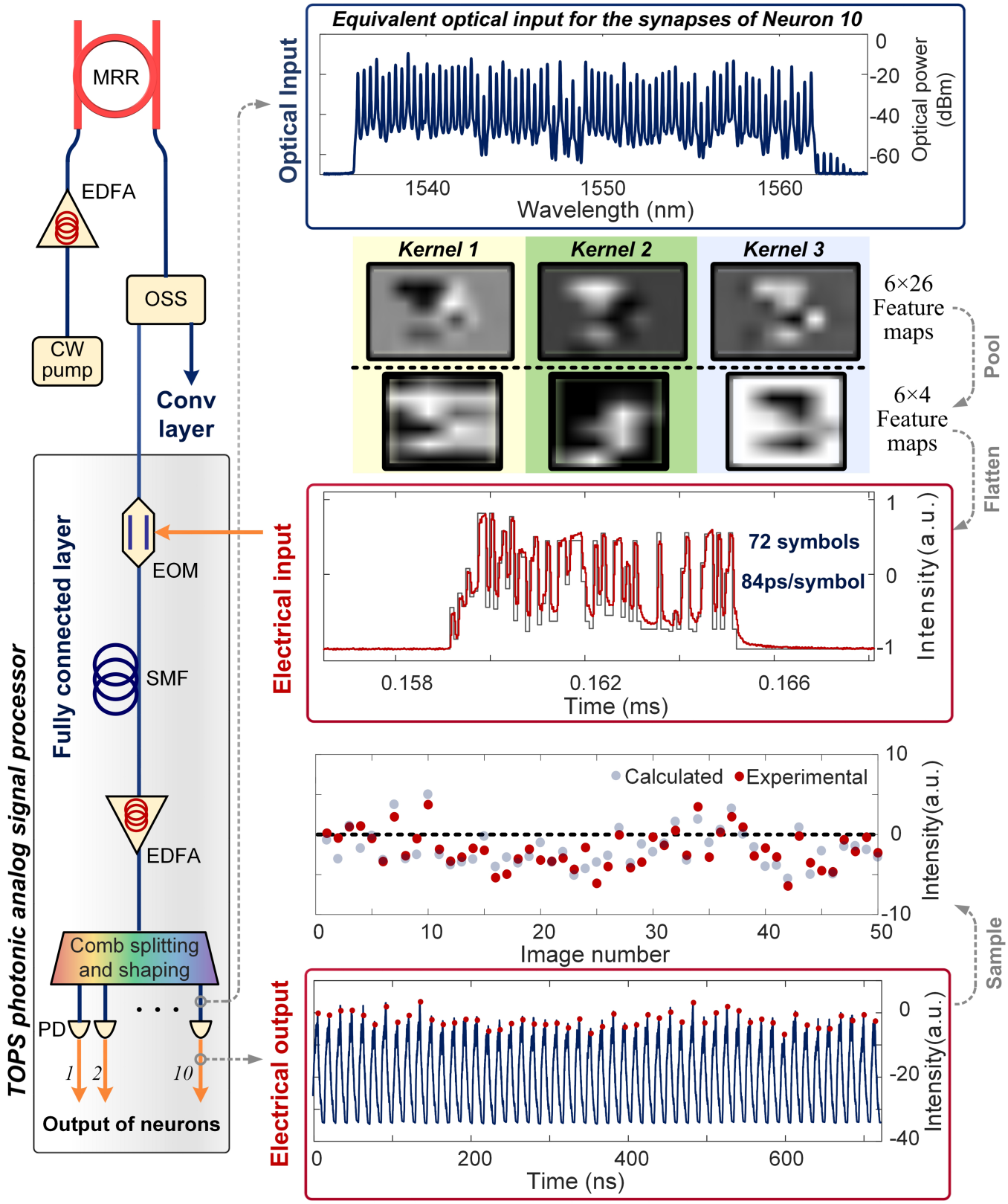
Extended Data Fig. 1 | VCA, for processing one-dimensional data. It consists of the experimental setup (right panel), the optical and electronic control and signal flow (left panel). ADC, analogue-to-digital converter. 1D, one-dimensional.



Extended Data Fig. 2 | Generation of soliton crystal microcombs. **a**, Schematic diagram of the soliton crystal microcomb, generated by pumping an on-chip high-*Q* (quality factor >1 million) nonlinear micro-ring resonator with a continuous-wave laser. **b**, Image of the MRR (upper inset) and a scanning electron microscope image of the MRR's waveguide cross-section (lower inset). **c**, Measured dispersion D_{int} of the MRR showing the mode crossing at about 1,552 nm. **d**, Measured soliton crystal step of the intra-cavity power. **e**, Optical spectrum of the microcomb when sweeping the pump wavelength. **f**, Optical spectrum of the generated coherent microcomb at different pump detunings at a fixed power. FSR, free spectral range.



Extended Data Fig. 3 | The architecture of the optical CNN. The architecture includes a convolutional layer, a pooling layer and a fully connected layer.



Extended Data Fig. 4 | Fully connected layers. Architecture and experimental results. The left panel depicts the experimental setup, similar to the convolutional layer. The right panel shows the experimental results for one output neuron, including the shaped comb spectrum (top); the pooled feature maps of the digit 3 and the corresponding input electrical waveform (the grey and red lines illustrate the ideal and experimentally generated waveforms, respectively; middle); and the output waveform of the neuron and sampled intensities (bottom). Conv layer, convolutional layer. CW pump, continuous-wave pump laser.