Article

Parallel convolutional processing using an integrated photonic tensor core

🔴 Check for updates

J. Feldmann[1,8], N. Youngblood[2,7,8], M. Karpov[3,8], H. Gehring[1], X. Li[2], M. Stappers[1], M. Le Gallo[4], X. Fu[3], A. Lukashchuk[3], A. S. Raja[3], J. Liu[3], C. D. Wright[5], A. Sebastian[4✉], T. J. Kippenberg[3✉], W. H. P. Pernice[1,6✉] & H. Bhaskaran[2✉]

With the proliferation of ultrahigh-speed mobile networks and internet-connected devices, along with the rise of artificial intelligence (AI)[1], the world is generating exponentially increasing amounts of data that need to be processed in a fast and efficient way. Highly parallelized, fast and scalable hardware is therefore becoming progressively more important[2]. Here we demonstrate a computationally specific integrated photonic hardware accelerator (tensor core) that is capable of operating at speeds of trillions of multiply-accumulate operations per second (10^{12} MAC operations per second or tera-MACs per second). The tensor core can be considered as the optical analogue of an application-specific integrated circuit (ASIC). It achieves parallelized photonic in-memory computing using phase-change-material memory arrays and photonic chip-based optical frequency combs (soliton microcombs[3]). The computation is reduced to measuring the optical transmission of reconfigurable and non-resonant passive components and can operate at a bandwidth exceeding 14 gigahertz, limited only by the speed of the modulators and photodetectors. Given recent advances in hybrid integration of soliton microcombs at microwave line rates[3–5], ultralow-loss silicon nitride waveguides[6,7], and high-speed on-chip detectors and modulators, our approach provides a path towards full complementary metal–oxide–semiconductor (CMOS) wafer-scale integration of the photonic tensor core. Although we focus on convolutional processing, more generally our results indicate the potential of integrated photonics for parallel, fast, and efficient computational hardware in data-heavy AI applications such as autonomous driving, live video processing, and next-generation cloud computing services.

The increased demand for machine learning on very large datasets[2] and the growing offering of AI services on the cloud[8,9] has driven a resurgence in custom hardware designed to accelerate MAC computations—the fundamental mathematical element needed for matrix-vector multiplication (MVM) operations. Although various custom silicon computing hardware—that is, field-programmable gate arrays (FPGAs)[10], ASICs[11] and graphics processing units (GPUs)[12]—have been developed to improve computational throughput and efficiency, they still depend on the same underlying electronic components, which are fundamentally limited in both speed and energy by Joule heating, electromagnetic crosstalk and capacitance[13]. The last of these (capacitance) dominates energy consumption and limits the maximum operating speed in neural network hardware accelerators[14]. This is because, the movement of data (for example, trained network weights), rather than arithmetic operations, requires the charging and discharging of chip-level metal interconnects. Thus, improving the efficiency of logic gates at the device level provides diminutive returns in such

applications, if the flow of data during computation is not simultaneously addressed[15]. Even recent developments such as memristive crossbar arrays[16–19] to compute in the analogue domain, although promising, do not have the potential for parallelizing the MVM operations (except by physical replication of the elements of the matrix). Moreover, they are plagued by the same limitations of electronic addressing[20], with additional challenges in the manufacturing and implementation due to issues with device variability[21,22], cyclability[23] and drift[24,25].

Integrated photonics benefits from the modularity and scalable fabrication methods of integrated circuits, while having two key advantages over its electronic counterparts: (1) massively parallel data transfer through wavelength division multiplexing (WDM) in conjunction with multichannel sources (that is, optical frequency combs); and (2) extremely high data modulation speeds limited only by the bandwidth of on-chip optical modulators and photodetectors. These uniquely photonic advantages have led to the ubiquity of optical networks for information transfer and are at present revolutionizing data centre interconnects

[1]Institute of Physics, University of Münster, Münster, Germany. [2]Department of Materials, University of Oxford, Oxford, UK. [3]Laboratory of Photonics and Quantum Measurements, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. [4]IBM Research Europe, Rüschlikon, Switzerland. [5]Department of Engineering, University of Exeter, Exeter, UK. [6]Center for Soft Nanoscience, University of Münster, Münster, Germany. [7]Present address: Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA. [8]These authors contributed equally: J. Feldmann, N. Youngblood, M. Karpov. ✉e-mail: ASE@zurich.ibm.com; tobias.kippenberg@epfl.ch; wolfram.pernice@uni-muenster.de; harish.bhaskaran@materials.ox.ac.uk
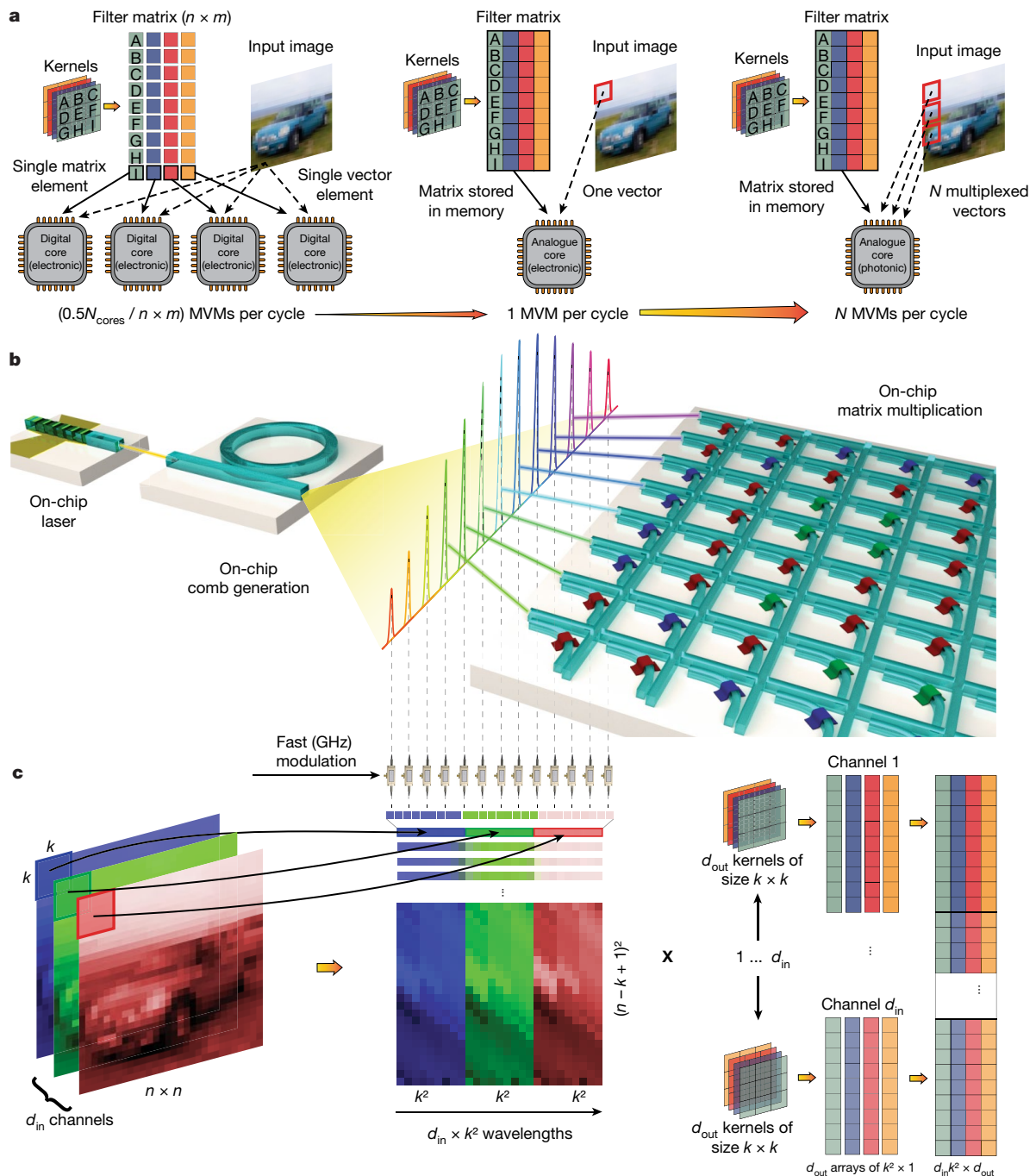
Fig. 1 | Photonic in-memory computing using a photonic-chip-based microcomb and PCMs. a, A comparison of digital and analogue electronic architectures with our photonic tensor core architecture. Digital electronics (left) requires many sequential processing steps distributed across multiple cores to compute convolutional operations on an image, whereas an entire MVM can be performed in one step using analogue electronic in-memory computing (centre). Photonic in-memory computing (right) brings wavelength multiplexing as an additional degree of freedom, enabling multiple MVM operations in a single time step. Photograph of car taken by author. **b**, Conceptual illustration of a fully integrated photonic architecture to compute convolutional operations. An on-chip laser (not used here) pumps an integrated Si_3N_4 microresonator to generate a broadband soliton frequency comb. Individual comb teeth, which form the input vectors, are modulated at high speeds, multiplied with a matrix of non-volatile phase-change memory cells, and summed along each column on a photodetector. **c**, An input image (left) with d_{in} channels is convolved with d_{out} kernels of size $k \times k$ by mapping convolution operations into a sequence of MVM operations. The input image is mapped to a series of $(n-k+1)^2$ input vectors of size $(d_{in} \times k^2) \times 1$ (middle) and multiplied by a filter matrix of dimension $(d_{in} \times k^2) \times d_{out}$ (right). Each comb line corresponds to one entry of the input vector and is modulated according to the pixel values of the input matrix.

(that is, server-to-switch communication). However, these developments have yet to seriously challenge digital electronics in the arena of information processing. Despite the current dominance of integrated electronics for computing, an application-specific optical processor not limited by the energy-bandwidth trade-off of electrical interconnects[13] could bring the advantages of optical networking to the field of computing. This would result in very high computational throughput via low-latency (that is, information processing and propagation at the speed of light) and parallel operations in a single physical optical processing core using WDM−essentially providing an additional scaling dimension through
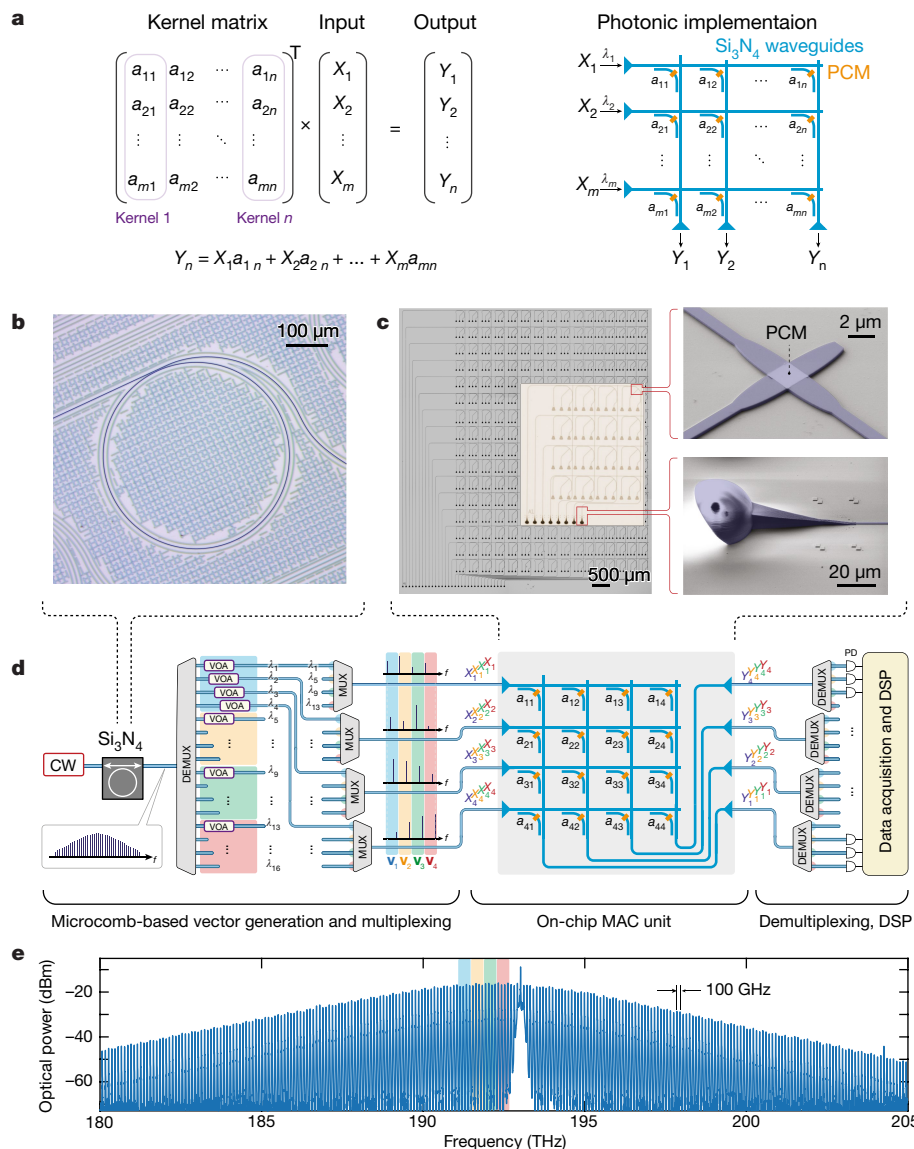
Fig. 2 | Concept of photonic tensor cores for convolution operations.
a, Basic MVM: a vector is encoded in the amplitude of individual comb teeth of a silicon nitride (Si_3N_4) photonic integrated soliton frequency comb (microcomb) exhibiting wavelengths ($\mathbf{X}_1$ to $\mathbf{X}_m$) and sent to the corresponding matrix input waveguides. The matrix elements are inscribed in the state of PCM patches on the waveguides. The splitting ratios of the directional couplers are chosen such that the same fraction of the light for each input reaches the output. **b**, Optical micrograph of a high-Q Si_3N_4 photonic-chip-based microresonator used for frequency comb generation. **c**, Optical micrograph of a fabricated 16 × 16. The inset shows a 4 × 4 matrix with 3D-printed input and output couplers to enable broadband operation. The close-up SEM images on the right show the 3D-printed couplers (botto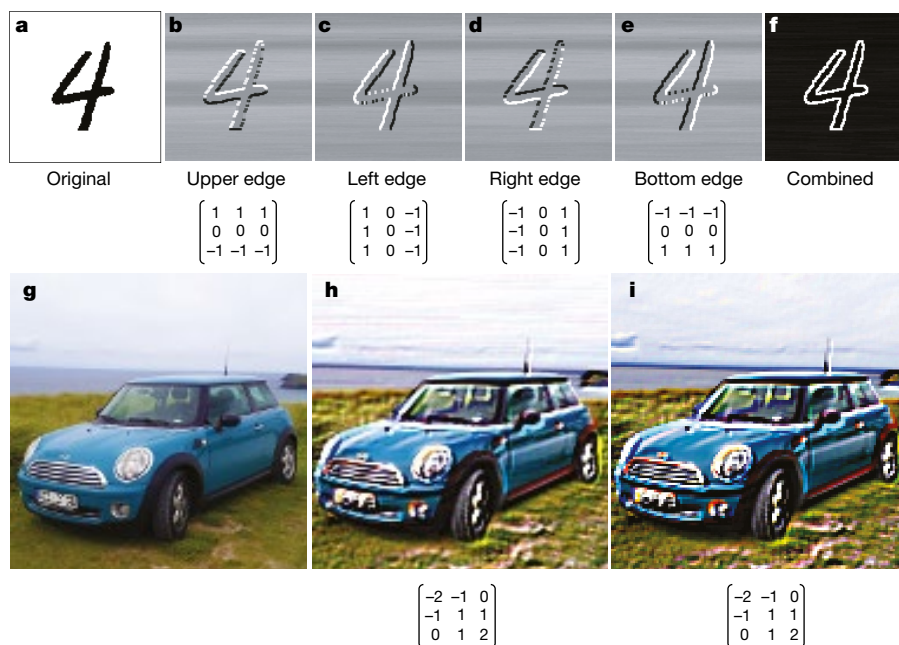m) and the waveguide crossings with the PCM (top) in more detail. **d**, Sketch of the multiplexed all-optical MVM. The input vectors are generated from lines of a photonic chip-scale DKS frequency comb driven by a continuous-wave (CW) laser using wavelength division multiplexers (MUXs) and variable optical attenuators (VOAs). The entries of different input vectors are grouped together again employing wavelength multiplexing and sent to the on-chip MAC unit that performs the calculations. After combining the correct wavelengths with optical wavelength division demultiplexers (DEMUXs), the multiplication results are obtained from the photodetectors (PD) followed by digital signal processing (DSP) as described in the main text. Note that in the given example four kernels and four input vectors are operated at once, resulting in 64 MAC operations per time step. **e**, Measured spectrum of a single-soliton frequency comb.

use of frequency space. Although the concept of free-space optics for efficient linear computing (for example, Fourier transforms, convolutions, matrix multiplication and so on) has existed for many decades[26] and continues to inspire computing architectures[27–30], precise control of the optical phase over the entire system remains the primary factor limiting scalability and commercialization of free space approaches.

Integrated photonics has the potential to solve these challenges. However, integration together with CMOS-compatible manufacturing is of paramount importance: on chip, both energy-efficient optical memory units and a compact, broadband multi-channel laser source must be combined within a scalable photonic architecture. Recent work on integrated photonic processors for MVMs and neuromorphic computing[31–33] has revealed the potential advantages of the photonic approach, but key issues such as large footprints (11,000 μm^2 per interferometer unit[31]) and the use of thermo-optic heaters to tune the phase or resonance wavelength of their components (ranging on average from 1 mW to 10 mW per heater for ring resonators and Mach–Zehnder interferometers, respectively) have been bottlenecks[34], as have devices such as add-drop resonators that limit the modulation bandwidth. Additionally, although using WDM for processing multiple inputs simultaneously in the same physical hardware has been proposed[35], it has not yet been demonstrated on-chip.

Original — Upper edge — Left edge — Right edge — Bottom edge — Combined

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$
$$\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$
$$\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} -2 & -1 & 0 \\ -1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$
$$\begin{bmatrix} -2 & -1 & 0 \\ -1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

Fig. 3 | Convolution using sequential MVM operations. a–e, Experimental result of convolving an image of 128×128 pixels showing a handwritten digit (**a**) with four image kernels of size 3×3 (corresponding to a 9×4 filter matrix). The kernels are chosen to highlight different edges of the input image. **f**, Combined image from **b–e** showing edge highlighting. **g–i**, Convolutional operation with a 3×3-sized image kernel (that is, emboss filter) without post-processing. The image in **g** shows the original image, whereas the other images depict the experimental (**h**) and the calculated (correct) (**i**) result. Photograph of car taken by author.

Here we design and experimentally demonstrate a scalable, CMOS-compatible, photonic hardware accelerator (which we term a 'photonic tensor core' in the following) capable of many parallel MVM operations at optical data rates to process images using convolutional filters (here, edge detection and emboss filters) and test it on the MNIST database[36] with a small-scale convolutional neural network (CNN). In a departure from electronic accelerators (see Fig. 1a), our photonic processor implements an on-chip matrix multiplication engine capable of performing parallel MAC operations using multiple wavelengths derived from a photonic chip-based optical frequency comb, which are incoherently added within a network of waveguides that exploit phase-change materials (PCMs). We leverage recent advances in chip-scale microcombs[3,4] operating in the regime of dissipative Kerr soliton (DKS) states, which generate broadband, low-noise and fully integrated optical frequency combs. Advances in low-loss Si_3N_4 photonic circuits based on the photonic Damascene process[6] have enabled microcomb line spacing in the microwave range compatible with direct electronic detection and power levels compatible with on-chip lasers[5,37,38]. Microcombs have already been employed in system-level demonstrations such as massively parallel coherent communications[39], optical frequency synthesizers[40] and massively parallel light detection and ranging (LiDAR)[41]. Thus far, DKS systems have, however, remained unexplored for photonic computing.

Key to our approach is the encoding of image data onto the individual comb teeth of an on-chip frequency comb, and subsequently encoding fixed convolutional kernels in the non-volatile configuration (that is, the amorphous or crystalline phase) of integrated PCM cells that couple evanescently to a matrix of interconnected photonic waveguides (shown in Fig. 1b). Our approach minimizes both latency and the movement of data, by using non-volatile in-memory photonic MAC operations and greatly reduces the footprint cost of photonics by multiplexing computations in the same photonic core. Importantly, both the soliton microcombs and the matrix of photonic waveguides can be implemented in silicon nitride[42], an ultralow-loss, CMOS-compatible nonlinear integrated photonic platform that is compatible with wafer-scale manufacturing and foundry. Combined with recent advances in both on-chip modulators and hybrid integration of soliton microcombs[5,37], fully integrated custom photonic tensor cores are viable.

Parallel 2D convolutions via MVM operations

One prominent class of machine learning models that benefit in terms of performance (speed, energy consumption) from high-throughput accelerators are CNNs, which are highly effective for applications such as in-image classification, autonomous navigation and audio analysis in the frequency domain. In state-of-the-art CNNs, many convolutional 'hidden layers' are applied to an input signal before feeding the processed data to fully connected layers for classification[43,44]. Each of the convolutional layers takes in an input image, performs convolutional operations to extract features and generates an output image. When performing convolutional operations in the digital domain, a minimum of two clock cycles are required for each sequential MAC operation—although the number of clock cycles for floating point multiplication usually exceeds three[45]. This leads to a substantial computational bottleneck, requiring distribution across multiple computing cores, as illustrated in Fig. 1a.

To build efficient hardware to perform the convolutional operations, one approach (originally conceived for electronic in-memory computing using memristive crossbar arrays[46,47]) is to combine all the convolutional filters into a large filter matrix stored in memory. As depicted in Fig. 1c, the filter matrix will be of dimension $(k^2 \times d_{in}) \times d_{out}$. It is constructed by stacking the kernel matrices into the columns of the final filter matrix. In the same way, the pixels of the input image are rearranged by stacking the pixels of the filter volume ($k \times k \times d_{in}$) into the rows of the input matrix. Hence, a single convolution operation involves $(n - k + 1)^2$ MVM operations between the filter matrix and the input vectors of dimension $k^2 \times d_{in}$. In the electronic domain, these MVM operations are typically multiplexed in time (serial processing) with parallelization afforded only by physically replicating the filter matrix. In this work,

Article
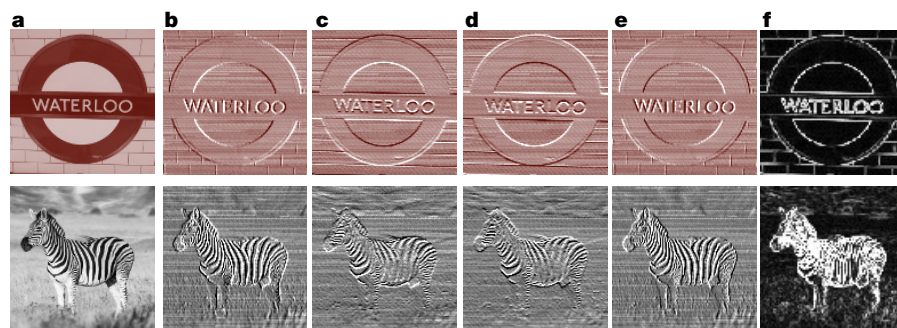
Fig. 4 | Convolution using parallel MVM operations. a–e, The original input images are shown in **a** (London Underground sign photograph taken by author; zebra photograph, WWU, Ehrman Photographic/ Shutterstock.com) and the output images using four different image kernels for highlighting edges are shown in **b–e**. The size of the four image kernels is 2×2 corresponding to a 4×4 filter matrix. In each time step, four input vectors are processed simultaneously via wavelength division multiplexing as illustrated in Fig. 2c. **f,** Combined image from **b–e** showing successful edge highlighting.

we exploit a photonic integrated soliton microcomb and optical WDM to overcome this fundamental limitation by encoding multiple input vectors of dimension $k^2 \times d_{in}$ onto multiple lines of a coherent chip-scale frequency comb. These optical input vectors can then be applied to a single $(k^2 \times d_{in}) \times d_{out}$ filter matrix simultaneously, thus eliminating duplicated physical hardware and sequential operations. This approach will be employed when designing the photonic tensor core.

The photonic tensor core

First, we demonstrate how to perform an MVM operation in the optical domain using photonic integrated circuits employing non-volatile PCM cells that store analogue values of the matrix in situ[48]. Details of using PCMs on single devices are described elsewhere[48,49]. In this work, the PCM ($Ge_2Sb_2Te_5$) cells are employed as attenuating matrix elements that absorb a desired amount of light depending on their particular phase configuration. In the crystalline PCM state, most of the incoming light is absorbed, representing for example a '0'. In the amorphous state, most of the light is transmitted, thus representing a '1'. Intermediate transmission states can be chosen by controllably switching fractions of amorphous and crystalline parts in the PCM cell[48,50]. To achieve both positive and negative matrix elements, we define '0' as an intermediate state between the crystalline and amorphous states, as described in the Supplementary Information.

To calculate the $m \times n$ MVM operation shown at the top of Fig. 2a, the input vector is encoded in the amplitude of the optical signals sent to the different matrix inputs. In addition to amplitude at a given wavelength, the input vector is also encoded at different wavelengths, thus enabling multiple calculations to be carried out simultaneously, while avoiding unwanted interference at the photodetector array.

Figure 2b depicts a scanning electron micrograph of the resonator used for comb generation and Fig. 2c shows an optical image of a fabricated 16×16 matrix with a 4×4 matrix as an inset. Key chip regions are magnified in the scanning electron micrographs on the right. Coupling of light into the optical chip is achieved using total internal reflection couplers[51,52] (lower inset of Fig. 2c), that allow the use of a wide wavelength spectrum. The PCM cells acting as the matrix elements are deposited on top of waveguide crossings (upper inset of Fig. 2c).

In addition to substantial benefits in modulation speed (for changing the vector inputs), an optical implementation of a matrix-vector multiplier allows the harnessing of wavelength division multiplexing to execute parallel MVM operations. In particular, as Fig. 2d shows, the same matrix can be used to process several input vectors at the same time when all the individual vectors are encoded in different wavelengths. Depending on the number of lines available in the frequency comb, the multiplexing scheme can be extended further, leading to substantial speed gains. Figure 2e shows the optical spectrum of an on-chip microcomb, revealing lines with 100-GHz spacing over a range of more than 25 THz.

To illustrate the principle outlined above experimentally, the convolution of an input image depicting a handwritten '4' (Fig. 3a) is performed using four 3×3 image kernels (resulting in a 9×4 filter matrix) and a single vector (9×1) per time step (Fig. 3b–e). Note that $d_{in} = 1$ and

$d_{out} = 4$ in this example. The image kernels applied in this example are chosen for edge detection and are shown below the output images (for how exactly the matrix elements are defined in the PCM state, see Supplementary Information section 8). After obtaining the results of the MVMs, the output values are offset by +0.5 and the values below 0 are set to 0 (black pixel) and the values above 1 are set to 1 (white). Each of the kernels highlights different edges of the original image: Fig. 3b, for example, highlights upper edges, whereas Fig. 3d brings out the opposite lower edges. Figure 3f shows the combined images (difference between alternating edges and addition of the two resulting images), highlighting that all edges have been properly detected. Since the four kernels are all inscribed in the same matrix, the pixel values of all four output images are obtained simultaneously, including more than 63,000 inner-product operations in total. The edge features are clearly visible, which emphasizes the effectiveness of our optical convolution operation. The inner product of the entire convolution was processed at approximately 1 kHz, limited only by the speed of the variable optical attenuators. Thus, owing to the slow electronic control and serial communication between the computer and microcontroller, the overall processing in this particular example took about four minutes. It should also be noted that in the examples of Fig. 3b–e, for each optical MVM, a software MVM operation is performed in a post-processing step to subtract a certain reference power from the measured output power in the matrix columns (more details are provided in the Supplementary Information).

To avoid the need for the above post-processing, the reference convolution operation can also be performed optically in the same on-chip matrix. In this case, one matrix column is in a reference state (see Supplementary Information). The output value from this column is then subtracted from all the matrix columns holding the actual image kernels. Figure 3g–i shows an experimental example of a convolution operation, which was performed without electrical post-processing using reference subtraction. Here, a 3×3 kernel (emboss filter) was applied using a 9×2 matrix, with one column for the image kernel and one column for the reference. The original image is shown on the left, whereas the experimental output image after the convolution operation is shown in the middle panel. From comparison with the calculated expected output on the right, it can be seen that the on-chip matrix performs well without the need for the post-processing step. We note that even though the image has three colour channels (red, green and blue; $d_{in} = 3$), the convolutions are performed on each channel independently and combined in the end, leading to the output image. However, this is more a limitation of the size of our hardware matrix than a fundamental limitation of this technology.

Having demonstrated the basic capabilities of our phase-change integrated photonic approach to performing convolutional operations in the optical domain, we now show, in Fig. 4, experimental examples of processing four input vectors in parallel at the same time. In this case, four pixels of the new image are obtained per image kernel simultaneously, thereby shortening the processing time by a factor of four. The kernel size used for this experiment is 2×2 and the input dimension of the image is $d_{in} = 1$, leading to a 4×4 filter matrix. The convolutions
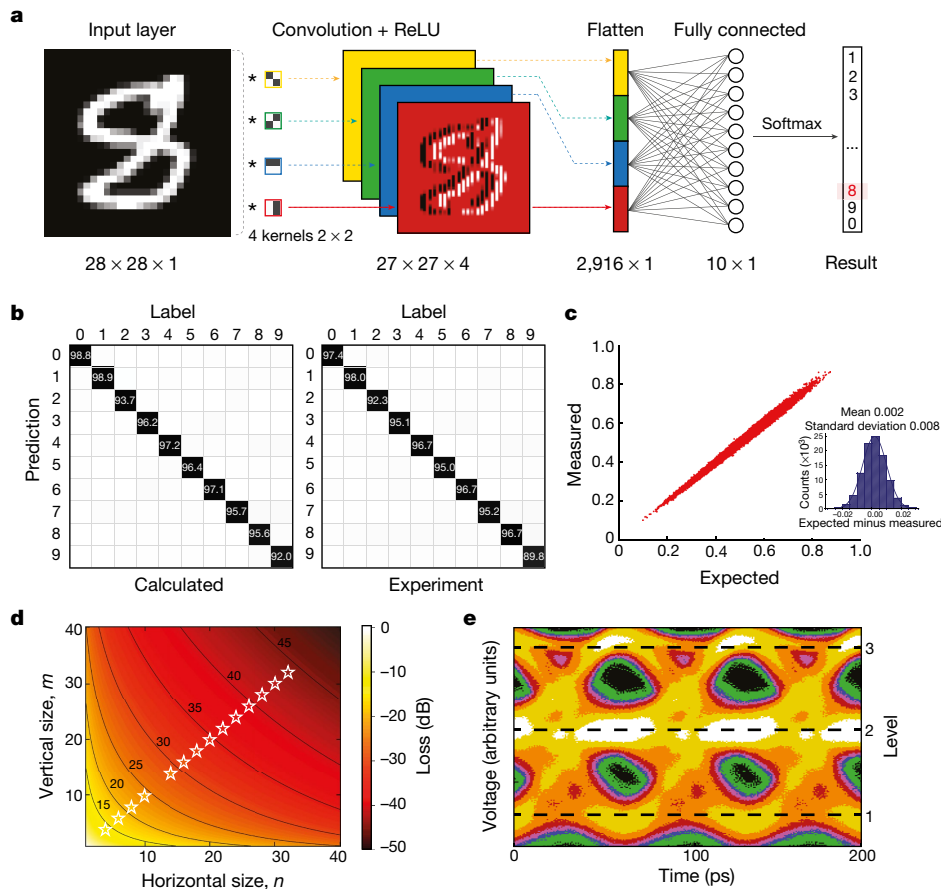
Fig. 5 | Digit recognition with a CNN and scalability. a, Layer structure of the network used to test the photonic tensor core with the MNIST handwritten digits database[36]. A rectified linear unit (ReLU) function is applied after the convolution and the Softmax function is applied in the classification layer. **b**, Confusion matrices showing similar performance for the prediction results for the experimental (95.3%) and calculated CNN (96.1%). **c**, Calculation accuracy for 100,000 MAC operations multiplying a vector of nine entries with a fixed matrix. The inset shows a histogram of the data revealing a standard deviation of 0.008 and therefore a resolution of 5 bits. **d**, Optical loss of the matrix as a function of its size. The heatmap depicts calculated optical loss for a directional coupler loss of 0.1 dB and a crossing loss of 0.12 dB. The stars represent measured optical loss for fabricated matrices. **e**, Eye-diagram for a matrix multiplication with a 2 × 1 matrix at a modulation speed of 13.5 GHz. The two inputs are modulated with two pseudo-random-bit-patterns, resulting in three different levels for the multiplication result.

again highlight different edges that can clearly be seen—for example, in the representation of the bricks in the upper image. Figure 4b and e emphasize vertical edges, whereas Fig. 4c and d highlight horizontal edges. This is in spite of variations in the vertical direction caused by power fluctuations of the input signal, underlining the robustness of the technique. Fig. 4f shows the combined images highlighting all edges. Given that four vectors are processed in parallel, the processing time is also decreased here by a factor of four compared to the results of Fig. 3.

Digit recognition with a CNN

Having shown that the photonic tensor core is capable of processing the convolutions demonstrated with different image filters, in a next step a CNN (see Fig. 5a) is built and tested against the MNIST handwritten digit database[36]. To test the accuracy of the predictions of the network, 10,000 test images were processed using the photonic matrix for the convolutions at a rate of 2 GHz (resulting in a processing time of 8.1 μs per image) with an FPGA for electronic control (more details on the experimental setup are given in the Supplementary Information and Supplementary Figs. 20–22). The confusion matrices illustrating the predictions for the different images with the experimentally obtained and the calculated results are shown in Fig. 5b. The experimental implementation of the CNN reached an accuracy of 95.3%, showing good agreement with the calculated prediction accuracy of 96.1%.

To analyse the computational accuracy of the optical convolutional processor for single dot-product operations, randomly chosen input vectors with nine entries are processed using a fixed matrix column and are compared against the expected analytically calculated multiplication result. The results for 100,000 calculations are scaled to the range [0, 1] and plotted in Fig. 5c together with the corresponding histogram, revealing a standard deviation of 0.008, which results in a resolution of 5 bits (more information on the evaluation of the resolution is given in the Supplementary Information and Supplementary Fig. 9).

Conclusion

We have described the first instance of a photonic tensor core that combines in-memory computing with state-of-the-art photonic integrated microcombs, enabling parallelizing convolution operations in the same physical device. We demonstrate simultaneous data transfer and computing at speeds comparable to fibre networks. Prior optical approaches to computing have largely been limited by a lack of integrated non-volatile photonic memory and the lack of multiplexing capability for such calculations[31,33,53]. Our approach overcomes both these limitations by (1) using non-volatile PCMs integrated onto waveguides to locally store convolutional kernels on-chip and (2) using photonic chip-based frequency combs to enable true in-memory photonic computing using WDM capability. The photonic tensor core demonstrated

Article

in this work is capable of operating at the speed of two tera-MAC operations per second (two trillion (10^{12}) MAC operations per second). Even faster operation, by an increase of several orders of magnitude, may be achievable by moderate scaling with state-of-the-art foundry processes. Using, for example, the smaller-footprint and industry-standard silicon-on-insulator platform, the matrix size can easily be scaled up to 40 × 40 (with acceptable loss; see Fig. 5d and Supplementary Fig. 10). With high modulation speeds exceeding 13 GHz (see the 2 × 1 MVM in Fig. 5e) available in the optical domain, computing densities of more than 400 TOPS per mm² with a throughput exceeding 1 peta-MAC operation per second (10^{15} MAC operations per second) can be achieved (see more details in the Methods and Supplementary Information, Supplementary Tables 1, 2 and Supplementary Figs. 3, 4).

A key feature of our approach is that, because the convolutional operation is a passive transmission measurement, the calculations can in theory be performed at the speed of light at very low power (17 fJ per MAC, considering only optical contributions), experimentally limited only by the modulation and detection bandwidths. Making use of the wavelength division multiplexing capabilities inherent to all-optical systems, our fast and parallelized implementation promises higher computational bandwidths than in electronic devices, because several pixels or even complete images can potentially be processed in a single time step. Our approach to convolutional processing provides an effective method for removing the computing bottleneck in machine learning hardware for applications ranging from live video processing to autonomous driving and AI-aided life-saving applications. More importantly, such an approach more broadly suggests that integrated photonics are coming of age and in some cases can begin to match and even challenge electronic computation.

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/s41586-020-03070-1.

1. Batra, G., Jacobson, Z., Madhav, S., Queirolo, A. & Santhanam, N. Artificial-intelligence hardware: new opportunities for semiconductor companies. https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies (McKinsey & Company, 2019).
2. Ben-Nun, T. & Hoefler, T. Demystifying parallel and distributed deep learning: an in-depth concurrency analysis. *ACM Comput. Surv.* **52**, https://doi.org/10.1145/3320060 (2019).
3. Herr, T. et al. Temporal solitons in optical microresonators. *Nat. Photon.* **8**, 145–152 (2014).
4. Herr, T., Gorodetsky, M. L. & Kippenberg, T. J. Dissipative Kerr solitons in optical microresonators. In *Nonlinear Optical Cavity Dynamics From Microresonators to Fiber Lasers* (ed. Grelu, P.) Vol. 8083, Ch. 6, 129–162 (Wiley, 2015).
5. Raja, A. S. et al. Electrically pumped photonic integrated soliton microcomb. *Nat. Commun.* **10**, 680 (2019).
6. Pfeiffer, M. H. P. et al. Photonic damascene process for integrated high-Q microresonator based nonlinear photonics. *Optica* **3**, 20–25 (2016).
7. Liu, J. et al. Ultralow-power chip-based soliton microcombs for photonic integration. *Optica* **5**, 1347–1353 (2019).
8. *Machine Learning on AWS* https://aws.amazon.com/machine-learning/ (accessed 12 October 2020).
9. Google Cloud *AI And Machine Learning Products* https://cloud.google.com/products/machine-learning/ (accessed 12 October 2020).
10. Zhang, C. et al. Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks. In *ACM/SIGDA Int. Symp. Field-Programmable Gate Arrays (FPGA '15)* https://doi.org/10.1145/2684746.2689060 (2015).
11. Jouppi, N. P. et al. In-datacenter performance analysis of a tensor processing unit. *Proc. ISCA '17* https://doi.org/10.1145/3079856.3080246 (2017).
12. Wang, P. S., Liu, Y., Guo, Y. X., Sun, C. Y. & Tong, X. O-CNN: octree-based convolutional neural networks for 3D shape analysis. *ACM Trans. Graph.* **36**, https://doi.org/10.1145/3072959.3073608 (2017).
13. Miller, D. A. B. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
14. Agrawal, S. R. et al. A many-core architecture for in-memory data processing. In *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO-50 '17)* 245–258, https://doi.org/10.1145/3123939.3123985 (IEEE/ACM, 2017).
15. Miller, D. A. B. Are optical transistors the logical next step? *Nat. Photon.* **4**, 3–5 (2010).
16. Ielmini, D. & Wong, H. S. P. In-memory computing with resistive switching devices. *Nat. Electron.* **1**, 333–343 (2018).
17. Le Gallo, M. et al. Mixed-precision in-memory computing. *Nat. Electron.* **1**, 246–253 (2018).
18. Boybat, I. et al. Neuromorphic computing with multi-memristive synapses. *Nat. Commun.* **9**, 2514 (2018).
19. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
20. Hu, M. et al. Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication. In *Proc. 53rd Annu. Design Automation Conf. (DAC '16)* https://doi.org/10.1145/2897937.2898010 (ACM Digital Library, 2016).
21. Gong, N. et al. Signal and noise extraction from analog memory elements for neuromorphic computing. *Nat. Commun.* **9**, 2102 (2018).
22. Joshi, V. et al. Accurate deep neural network inference using computational phase-change memory. *Nat. Commun.* **11**, 2473 (2020).
23. Yang, T. Y., Park, I. M., Kim, B. J. & Joo, Y. C. Atomic migration in molten and crystalline $Ge_2Sb_2Te_5$ under high electric field. *Appl. Phys. Lett.* **95**, 032104 (2009).
24. Koelmans, W. W. et al. Projected phase-change memory devices. *Nat. Commun.* **6**, 8181 (2015).
25. Kim, S. et al. A phase change memory cell with metallic surfactant layer as a resistance drift stabilizer. In *2013 IEEE Int. Electron Devices Meeting* https://doi.org/10.1109/IEDM.2013.6724727 (IEEE, 2013).
26. Bell, T. E. Optical computing: a field in flux: a worldwide race is on to develop machines that compute with photons instead of electrons but what is the best approach? *IEEE Spectr.* **23**, 34–38 (1986).
27. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M. & Englund, D. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2018).
28. Silva, A. et al. Performing mathematical operations with metamaterials. *Science* **343**, 160–163 (2014).
29. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
30. Colburn, S., Chu, Y., Shilzerman, E. & Majumdar, A. Optical frontend for a convolutional neural network. *Appl. Opt.* **58**, 3179–3186 (2019).
31. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
32. Tait, A. N. et al. Silicon photonic modulator neuron. *Phys. Rev. Appl.* **11**, 064043 (2019).
33. Pérez, D. et al. Multipurpose silicon photonics signal processor core. *Nat. Commun.* **8**, 636 (2017).
34. Galal, S. & Horowitz, M. Energy-efficient floating-point unit design. *IEEE Trans. Comput.* **60**, 913–922 (2011).
35. Bangari, V. et al. Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs). *IEEE J. Sel. Top. Quantum Electron.* **26**, https://doi.org/10.1109/JSTQE.2019.2945540 (2020).
36. LeCun, Y., Cortes, C. & Borges, C. J. C. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist.
37. Stern, B., Ji, X., Okawachi, Y., Gaeta, A. L. & Lipson, M. Battery-operated integrated frequency comb generator. *Nature* **562**, 401–405 (2018).
38. Jones, R. et al. Heterogeneously integrated InP/silicon photonics: fabricating fully functional transceivers. *IEEE Nanotechnol. Mag.* **13**, 17–26 (2019).
39. Marin-Palomo, P. et al. Microresonator-based solitons for massively parallel coherent optical communications. *Nature* **546**, 274–279 (2017).
40. Spencer, D. T. et al. An optical-frequency synthesizer using integrated photonics. *Nature* **557**, 81–85 (2018).
41. Riemensberger, J. et al. Massively parallel coherent laser ranging using soliton microcombs. *Nature* **581**, 164–170 (2019).
42. 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).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/CVPR.2016.90 (IEEE, 2016).
44. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd Int. Conf. Learning Representations (ICLR 2015)* (eds Bengio, Y. & LeCun, Y.) 4 (2015); https://arxiv.org/abs/1409.1556.
45. Al-Ashrafy, M., Salem, A. & Anis, W. An efficient implementation of floating point multiplier. In *2011 Saudi Int. Electronics, Communications and Photonics Conf. (SIECPC)* https://doi.org/10.1109/SIECPC.2011.5876905 (2011).
46. Gao, L., Chen, P. Y. & Yu, S. Demonstration of convolution kernel operation on resistive cross-point array. *IEEE Electron Device Lett.* **37**, 870–873 (2016).
47. Shafiee, A. et al. ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In *Proc. 2016 43rd Int. Symp. Computer Architecture (ISCA 2016)* https://doi.org/10.1109/ISCA.2016.12 (2016).
48. Li, X. et al. Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell. *Optica* **6**, 1–6 (2019).
49. Ríos, C. et al. Integrated all-photonic non-volatile multi-level memory. *Nat. Photon.* **9**, 725–732 (2015).
50. Feldmann, J. et al. Calculating with light using a chip-scale all-optical abacus. *Nat. Commun.* **8**, 1256 (2017).
51. Gehring, H. et al. Low-loss fiber-to-chip couplers with ultrawide optical bandwidth. *APL Photon.* **4**, 010801 (2019).
52. Gehring, H., Eich, A., Schuck, C. & Pernice, W. H. P. Broadband out-of-plane coupling at visible wavelengths. *Opt. Lett.* **44**, 5089 (2019).
53. Nahmias, M. A. et al. Photonic multiply-accumulate operations for neural networks. *IEEE J. Sel. Top. Quantum Electron.* https://doi.org/10.1109/jstqe.2019.2941485 (2019).

Methods

Device fabrication

The photonic circuits used for the convolution experiments are fabricated using a three-step electron-beam lithography (Raith EBPG 5150) process on a silicon nitride (325 nm) on silicon oxide (3,300 nm) on silicon wafer (Rogue Valley Microdevices). The complete circuit was designed using GDShelpers, a design framework for integrated circuitry[54].

In the first lithography step, windows in the positive tone resist (poly(methylmethacrylate), PMMA) are exposed for the deposition of alignment markers made from gold. The resist is developed in 1:3 methyl isobutyl ketone (MIBK):isopropanol for 120 s and a layer stack of 5 nm chromium, 120 nm gold and 5 nm chromium are evaporated via electron-beam physical vapour deposition. By sonicating the chip in acetone, the PMMA is removed and only the gold markers in the exposed positions remain. The markers are used in the second step to align the photonic structures. After spin-coating a layer of 300 nm of the resist and prebaking it for 60 s at 85 °C, an etch mask is exposed in the negative-tone electron-beam resist arN 7520.12 (Allresis). The photonic structures are developed in MF-319 (Allresis) for 75 s and a post-development bake is performed at 85 °C for 60 s. By using reactive ion etching with a CHF_3/O_2 plasma, the mask of the photonic circuits is transferred into the sample. The silicon nitride layer is fully etched leaving single mode waveguides at telecommunications wavelengths with a width of 1.2 μm and a height of 325 nm. Subsequently the remaining resist is removed in an oxygen plasma for 10 min. In the third electron-beam lithography step, windows for the deposition of the PCM are written using the same markers as for the photonic structures for the alignment. The same process as in the first electron-beam lithography step is used. Finally, 10 nm of the PCM ($Ge_2Sb_2Te_5$) and 10 nm of indium tin oxide (ITO) are sputter-deposited on the sample. Both layers are sputtered using radio-frequency sputtering with an argon plasma (5-mTorr pressure, 15 standard cubic centimeters per minute (sccm) Ar, 30-W radio-frequency power and a base pressure of 2×10^{-6} Torr). The ITO is used as a protective film to prevent oxidation of the PCM. As in the marker-deposition, the PMMA is lifted off by sonicating the sample in acetone, leaving the PCM only in the desired positions on the photonic circuitry. Prior to the experiments the $Ge_2Sb_2Te_5$ is crystallized on a hot plate at 220 °C for approximately 10 min.

Measurement setup

The experimental setups used to perform the convolution experiments are shown in Supplementary Figs. 1–3. The individual wavelengths are generated using a frequency comb that is operated in the single soliton state and separated using a fibre-based multiplexer. For the image processing experiments (Figs. 3 and 4) the wavelengths (input vectors) are modulated using variable optical attenuators based on micro-electro-mechanical systems, whereas the fast modulation (Fig. 5) was performed with a 20 GHz electro-optic modulator. The input signal is coupled to the chip using 3D-printed broadband total internal reflexion couplers (see Supplementary Figs. 17 and 18) capable of operating from the visible to the telecommunications-wavelength regime.

In the multiplexed version of the experiment, processing four vectors at the same time, the corresponding wavelengths are multiplexed and demultiplexed accordingly before and after the matrix, again using fibre-multiplexers. The convolution results are read using photodetectors (New Focus Model 2011). In the frequency response experiment (Fig. 5), a fast photodiode (12 GHz) was used.

The measurement setup remains stable for extended periods, as also detailed in the Supplementary Information (Supplementary Fig. 11). Fluctuations in the transmission are due to temperature variations in the laboratory, which oscillate during day and night times. The long-term trend, however, remains unchanged over weeks (see Supplementary Fig. 7) and also months[53].

Realization of high-Q Si$_3$N$_4$ microresonators

The soliton microcombs used in our work are based on Si_3N_4 micro-ring resonators with a free spectral range of 100 GHz shown in Fig. 2b. The micro-resonators are fabricated using the photonic damascene process[6], which provides access to high quality factors (Q factors, reaching 10^7) and enables the four-wave-mixing-based nonlinear frequency-conversion processes as well as the formation of DKS states at low pump powers[7].

The microresonators were designed to have cross-section dimensions of 0.82 μm × 1.50 μm, which ensure anomalous group velocity dispersion of about 1–2 MHz at around 1,550 nm, as needed for the Kerr comb generation and the formation of DKS states. The light is coupled evanescently to a microresonator via the on-chip bus waveguide (with similar dimensions) located close to the microring, and which are additionally equipped with inverse tapers at the ends for edge chip coupling. The Si_3N_4 chips we used are furthermore fibre-packaged with an average loss of 4 dB per interface to facilitate light coupling in and out of the system. The fabricated devices have Q-factors exceeding 5×10^6, which allows for DKS generation and switching[55,56] even for relatively low input pump powers below 1 W.

Soliton comb generation

For the DKS generation a Si_3N_4 microring resonator is driven using a continuous-wave tunable fibre laser which is amplified with an erbium-doped fibre amplifier (EDFA) to a power level of about 1 W. A high-power bandpass filter is used to suppress the amplified spontaneous emission from the EDFA. The light polarization is adjusted using a fibre-based polarization controller to match the transverse electric polarized fundamental mode of the microresonator, and then is launched to the fibre-coupled Si_3N_4 chip.

To launch the DKS state, a standard pump tuning technique is applied[3], in which the amplified seed laser is swept over the chosen frequency resonance from the blue-detuned side to the red-detuned side at a speed of approximately 200 GHz s^{-1}. This approach allows us to generate multiple-soliton states with several pulses inside the cavity, which, however, usually has a highly structured optical spectrum. To achieve the single DKS state with a spectrally smooth sech^2-shaped envelope the soliton switching procedure is employed[55] and the pump is slowly tuned towards shorter wavelengths until the single soliton state is stabilized. To improve the long-term stability of the generated DKS states and align the resulting optical frequency comb to the established International Telecommunication Union grids, the Si_3N_4 chip is thermally controlled, which enables the use of the standard WDM equipment and optical comb stabilization against environmental temperature fluctuations and setup drifts, ensuring >8 h of continuous operation.

The resulting DKS-based optical frequency comb with 100-GHz line spacing and spanning over multiple telecommunication bands is coupled out from the chip. The residual pump is suppressed using a fibre-based notch filter, and a small portion of the light (1%) is used for monitoring purposes. The rest of the comb is shown in Fig. 2e, and is then additionally amplified with C-band EDFA to further employ it in the setup for encoding and demultiplexing of the image vectors. The amplification of the EDFA of up to 15 dB was individually chosen for the different experiments and is mainly used to compensate for coupling losses between the fibre array and the chip.

Details of the convolution operation in a CNN

For the convolution between an input image of dimension $n \times n$ with d_{in} channels and a filter of dimension $k \times k \times d_{in}$, the resulting output image is of dimension $(n - k + 1) \times (n - k + 1)$. To perform each convolutional operation, a filter is passed over the input image, inspecting a small window of pixels at a time. A pixel-wise MAC operation between the filter and the current filter window is carried out to calculate a

single pixel of the output image. In CNNs, d_{out} convolutional kernels will be applied to the same image, which corresponds to $(n - k + 1)^2 \times k^2 \times d_{in} \times d_{out}$ MAC operations per convolution layer and scales in computational complexity as $O(n^2k^2)$. It is worth noting that for the case of large kernels ($k > 15$), performing the convolution in the Fourier domain can reduce computational complexity[57] to about $O(n^2\ln(n))$. However, $k \leq 5$ for most kernels in many common CNN models used today (that is, AlexNet[58], ResNet[43], GoogLeNet[59], and so on) making the Fourier approach less efficient than direct convolution.

Implementation of the photonic matrix

To perform an MVM each vector entry ($X_1, ... X_m$) is encoded on a separate wavelength (see Fig. 2a). Therefore, the input vectors can be fed to the matrix by modulating the input signals with currently available fast electro-optical modulators, providing access to very high data rates. The matrix itself is designed as a waveguide crossbar array with additional directional couplers that equally distribute the input power to all PCM cells (more details of the splitting ratios of the directional couplers are given in the Supplementary Information and Supplementary Figs. 7, 8). The matrix elements are encoded in the state of the PCM and programmed optically through additional inputs in each matrix cell (Supplementary Fig. 5). By using a soliton microcomb with a mode spacing that exceeds the detector bandwidth, interference inside the waveguides can be avoided and the summation of the individual products (of the MVMs) can be performed by adding the comb teeth to the output waveguides, also by using directional couplers. With the horizontal directional couplers, the input vectors are equally distributed to the different columns of the matrix (which represent the individual image kernels), whereas the vertical directional couplers combine the input light after interaction with the PCM cells and perform the accumulation operation. It should be noted that each vector entry interacts only with a single PCM cell per matrix column. This interaction can be viewed as a single multiplication between the incoming amplitude and the absorption of the PCM cell, as has been shown in previous work[60]. The output power at each column of the matrix represents the inner-product (the sum of the individual products) of the input vector with a kernel multiplied by a certain (fixed) factor of $1/(m \times n)$, which depends on the matrix size. Power distribution due to fan-out accounts for the $1/n$ loss, whereas combining m non-interfering sources with directional couplers accounts for the additional $1/m$ loss due to energy conservation.

Parallel MVMs

To increase the compute density, the throughput of the photonic tensor core multiple vectors can be fed to the matrix at the same time, making use of wavelength division multiplexing (as shown in Fig. 2d). The wavelengths needed to encode the vectors are generated using a single DKS state of a microcomb[4,6,61] which is fed into a demultiplexer to split up the individual wavelengths (λ_1 to λ_{16}). After manipulating the amplitude of each comb line individually (according to the value of the input vectors) by using variable optical attenuators, the corresponding entries of each vector are multiplexed back together (that is, λ_1, $\lambda_5, \lambda_9, \lambda_{13}$) and sent to the matrix input. After propagating through the filter matrix, all output waveguides of the matrix contain all 16 input wavelengths. Proper demultiplexing and combining of the wavelengths corresponding to the individual vectors yields the convolutional results that can be measured with photodetectors. In the current example, 16 inner-product operations (four kernels applied to four input vectors) are carried out in a single time step.

Details of the accuracy measurements

The accuracy measurement of the experiment shown in Fig. 5c was carried out using the same setup as for edge detection, using variable optical attenuators operated in the kilohertz regime and a 9 × 1 matrix. We note, however, that the main source limiting the precision of MVMs

in our architecture are the electronic signals driving the modulators and the extinction ratio of the modulator (both of which do not depend on frequency). Therefore, no loss in precision is expected when driving the system at higher speeds. We also note that the matrix elements can be programmed with a precision of more than 8 bits using a closed-loop approach, as shown in Supplementary Fig. 6.

Structure and implementation of the CNN

The CNN employed in our experiments is depicted in Fig. 5a and consists of the input layer taking the pixel data (28 × 28 pixels, single channel) that is then passed to a convolutional layer consisting of four 2 × 2 kernels plus subsequent Rectified Linear Unit (ReLU) activation, resulting in an output of dimension 27 × 27 × 4 (valid padding). The output from the convolution step is flattened and fed to a fully connected layer with ten neurons. The probabilities for every digit are obtained from the final classification using the Softmax function. The network was trained via software (see Supplementary Information section 14 for more detail) and the weights of the filter kernels were programmed to the states of the PCM cells in the on-chip matrix.

Projections to the future

The experimental data in the main paper were obtained with matrices up to a size of 9 × 4, with a maximum of four input vectors per time step and a modulation speed of up to 2 GHz. To estimate the ultimate performance capabilities of the system, we now explore the scaling capabilities in terms of matrix size, modulation speed and number of parallel vectors. The main factor affecting the achievable matrix size is the optical loss induced by the photonic matrix, which results from equally splitting the light to all matrix cells, combined with the insertion loss of the directional couplers and waveguide crossings. As detailed in the description of the construction of the photonic tensor core (detailed in the Supplementary Information), the optical loss in the matrix itself scales with the matrix size as $1/(m \times n)$ for a matrix size $m \times n$. Additional loss is added by the directional couplers and the waveguide crossings and increases linearly with the matrix size. The propagation loss of the waveguides (0.2 dB cm^{-1}) can be neglected in comparison to these contributions. Figure 5d shows a heatmap of the calculated matrix loss as a function of the matrix size, considering measured insertion loss of 0.1 dB per directional coupler and 0.12 dB per crossing (see Supplementary Information sections 4 and 6). The stars on the diagonal represent measured optical loss for fabricated matrices with sizes up to 32 × 32, and agree well with the calculations (see Supplementary Figs. 10 and 15). By further improving the loss of the crossings[62] and directional couplers the limits to matrix sizes can be increased.

To illustrate convolutional processing using high-speed modulation of the input vectors, Fig. 5e shows an eye diagram at a modulation speed of 13.5 GHz obtained from a 2 × 1 matrix. The two electro-optical input modulators (bandwidth 40 GHz) were driven by 2^7 pseudo-random-bit patterns provided by a fast pulse-generator, thus resulting in three output levels that can be clearly distinguished. As the photonic matrix itself is operated passively in a transmission measurement, the speed is limited only by the bandwidth of the modulators and detectors. In the experiment, a detector with a 3-dB bandwidth of 12 GHz was used (additional data on modulating the individual matrix inputs up to 14 GHz were included in the Supplementary Information).

Because the photonic system is designed with broadband input couplers and broadband directional couplers in silicon nitride with a wide optical transparency window, the tensor processor supports more than 200 individual wavelengths from the frequency comb source with a spacing of 100 GHz (see Supplementary Information section 13 and Supplementary Fig. 19). In addition to the spectral width of the frequency comb, the influence of wavelength-dependent parts in the matrix design must also be considered when estimating the wavelength range exploitable for the calculations. In this case, it is predominantly

the wavelength dependence of the directional couplers that hinders the equal distribution of the input power for all wavelengths. Whereas our design offers an impressive range of approximately 100 nm, this can be considerably improved by an adapted design[63]. The influence of dispersion in the PCM absorption can be neglected in the wavelength range considered and could be corrected by adjusting the input amplitudes of the different comb lines. Thus, for a 9×4 matrix, four multiplexed input vectors and a modulation speed of 14 GHz, a processing speed of 2 trillions (10^{12}) of MAC operations per second (9×4 MACs $\times 4$ input vectors $\times 14$ GHz) can be obtained. This, however, is not the ultimate speed, since we are limited here by the modulation and detection bandwidth of our particular experimental setup.

When comparing optical architectures with digital electronics, it is helpful to use compute density (defined here as TOPS (trillions of operations per second) normalized by the processor area[53]) as a figure-of-merit for performance. This helps us to directly compare the processing throughput of architectures that may employ very different schemes for computing MVM operations. For the SiN devices demonstrated here, the area of a single MAC (with one MAC being two operations) unit cell is $285\,\mu m \times 354\,\mu m$. This, when operating at 12 GHz with 4 input vectors via WDM, corresponds to a compute density of 1.2 TOPS per mm^2. By moving to a silicon-on-insulator platform with a nominal bend radius of $5\,\mu m$ and using integrated electrical control of the $Ge_2Sb_2Te_5$ (refs. [64,65]), it would be straight-forward to reduce the area of the MAC unit cell to less than $30 \times 30\,\mu m^2$, increasing the compute density to 420 TOPS per mm^2 per input channel (see Supplementary Information section 11. We also demonstrate increased compute density with a silicon-on-insulator prototype illustrating the feasibility of this approach (Supplementary Figs. 12–14) and scaling linearly with the number of input vectors via WDM—a notably different computing paradigm compared to electronic approaches (note that the compute density considers only the photonic tensor core itself, without the electronic control and the off-chip multiplexers). The energy efficiency for the actual experiments can be calculated to be 0.4 TOPS per Watt for 5-bit resolution (including the optical power as well as the analogue-to-digital converters and modulators which we estimate to be dominating the consumption; see Supplementary Information). Moreover, by reducing the loss of the directional couplers and waveguide crossings and integrating detectors and modulators on-chip, the efficiency can be increased to 7.0 TOPS per Watt in the future. Considering only the optical energy based on the power needed to overcome shot-noise for a fixed 8-bit precision number at the output, the energy per MAC operation can be as low as 17 fJ per MAC (see Supplementary Fig. 16).

To estimate the full capabilities of the optical accelerator for convolutional operations, the performance of common optical components in foundry services[66,67] must be considered in combination with the wavelength range of the frequency comb that can be used. The frequency comb clearly shows lines from 1,500 nm to 1,650 nm (see Supplementary Information), leading to a range of 150 nm exploitable for computation that can be extended by optimizing the setup. Considering the spacing of the comb lines of 100 GHz (0.8 nm), this leads to approximately 150 nm/0.8 nm = 187 different wavelengths. Decreasing the spacing to 50 GHz (0.4 nm) and increasing the matrix size to 50×50, the operational speed can reach an unprecedented 1 peta-MAC operations per second (that is, a quadrillion (10^{15}) MAC operations per second) for a single matrix, assuming a modulation and detection speed of 50 GHz. These large matrix sizes are experimentally feasible using variable-length directional couplers and have been demonstrated using a photonics foundry process in 2013[68].

Data availability
All data used in this study are available from the corresponding author upon reasonable request.

54. Gehring, H., Blaicher, M., Hartmann, W. & Pernice, W. H. P. Python based open source design framework for integrated nanophotonic and superconducting circuitry with 2D-3D-hybrid integration. *OSA Continuum* **2**, 3091–3101 (2019).
55. Guo, H. et al. Universal dynamics and deterministic switching of dissipative Kerr solitons in optical microresonators. *Nat. Phys.* **13**, 94–102 (2017).
56. Karpov, M. et al. Dynamics of soliton crystals in optical microresonators. *Nat. Phys.* **15**, 1071–1077 (2019).
57. Fialka, O. & Čadík, M. FFT and convolution performance in image filtering on GPU. In *Proc. 10th Int. Conf. Information Visualisation (IV'06)* https://doi.org/10.1109/IV.2006.53 (IEEE, 2006).
58. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, https://doi.org/10.1145/3065386 (2017).
59. Szegedy, C. et al. Going deeper with convolutions. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/CVPR.2015.7298594 (IEEE, 2015).
60. Ríos, C. et al. In-memory computing on a photonic platform. *Sci. Adv.* **5**, eaau5759 (2019).
61. Gaeta, A. L., Lipson, M. & Kippenberg, T. J. Photonic-chip-based frequency combs. *Nat. Photon.* **13**, 158–169 (2019).
62. Ma, Y. et al. Ultralow loss single layer submicron silicon waveguide crossing for SOI optical interconnect. *Opt. Express* **21**, 29374–29382 (2013).
63. Lu, Z. et al. Broadband silicon photonic directional coupler using asymmetric-waveguide based phase control. *Opt. Express* **23**, 3795–3808 (2015).
64. Farmakidis, N. et al. Plasmonic nanogap enhanced phase change devices with dual electrical-optical functionality. *Sci. Adv.* **5**, eaaw2687 (2019).
65. Zhang, H. et al. Miniature multilevel optical memristive switch using phase change material. *ACS Photon.* **6**, 2205–2212 (2019).
66. Atabaki, A. H. et al. Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip. *Nature* **556**, 349–354 (2018).
67. Wang, X. & Liu, J. Emerging technologies in Si active photonics. *J. Semicond.* **39**, 061001 (2018).
68. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S. & Watts, M. R. Large-scale nanophotonic phased array. *Nature* **493**, 195–199 (2013).

Author contributions W.H.P.P., H.B., A.S., T.J.K. and C.D.W. conceived the experiment. J.F. fabricated the devices with assistance from N.Y., H.G. and X.L. N.Y. performed the deposition of the $Ge_2Sb_2Te_5$ material, together with X.L. J.F. implemented the measurement setup and carried out the measurements with help from N.Y., M.K., M.S. and H.G. M.K., X.F., A.L., A.S.R. and J.L. implemented the frequency comb source. All authors discussed the data and wrote the manuscript together.