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].

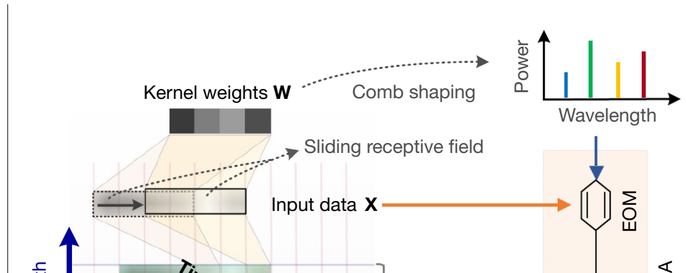
The capability of neural networks is dictated by the computing power of the underlying neuromorphic hardware. Optical neural networks (ONNs)[8–12] are promising candidates for next-generation neuromorphic computation, because they have the potential to overcome some of the bandwidth bottlenecks of their 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

10.1038/s41586-020-03063-0
View Article Page



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