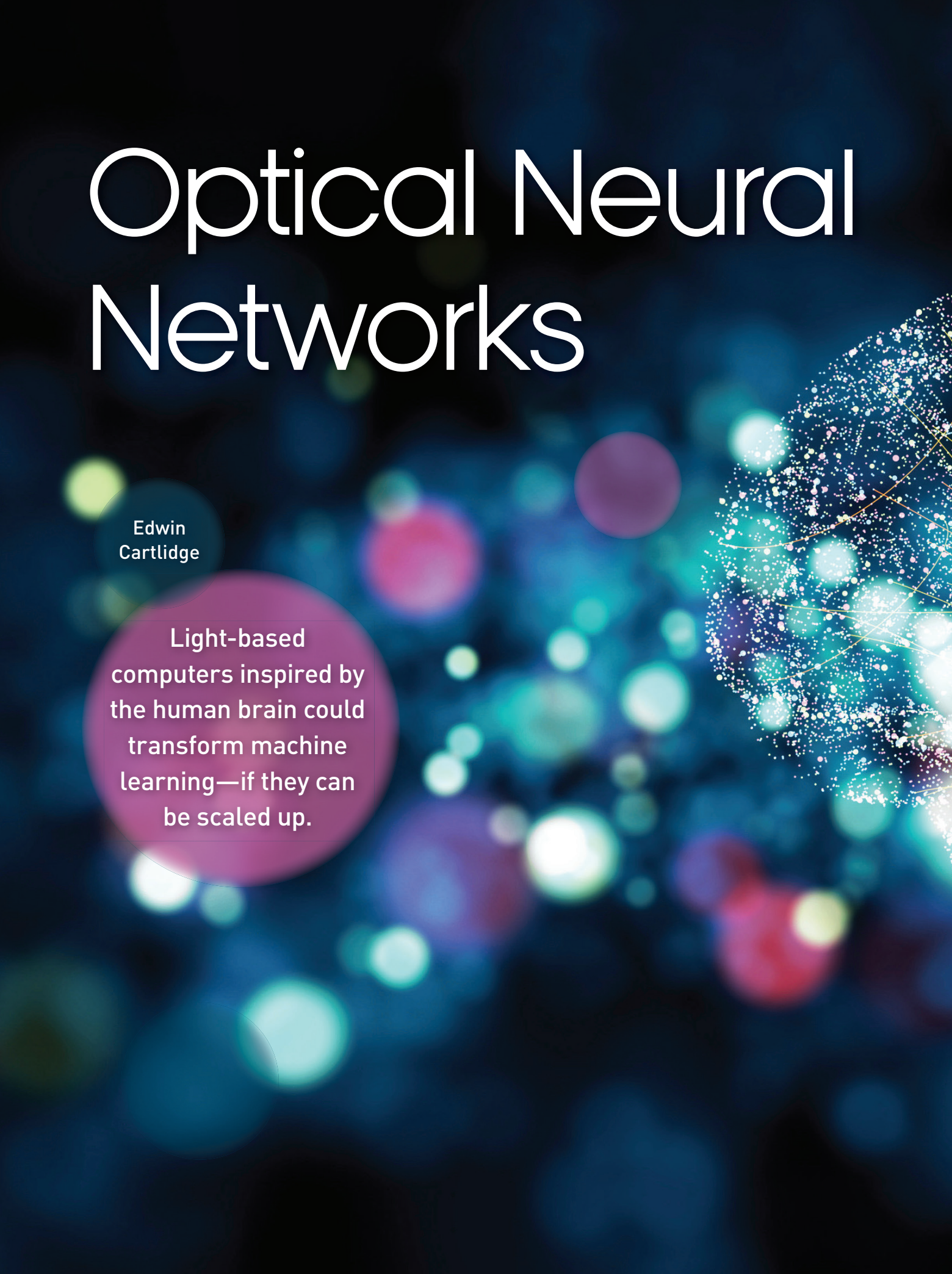
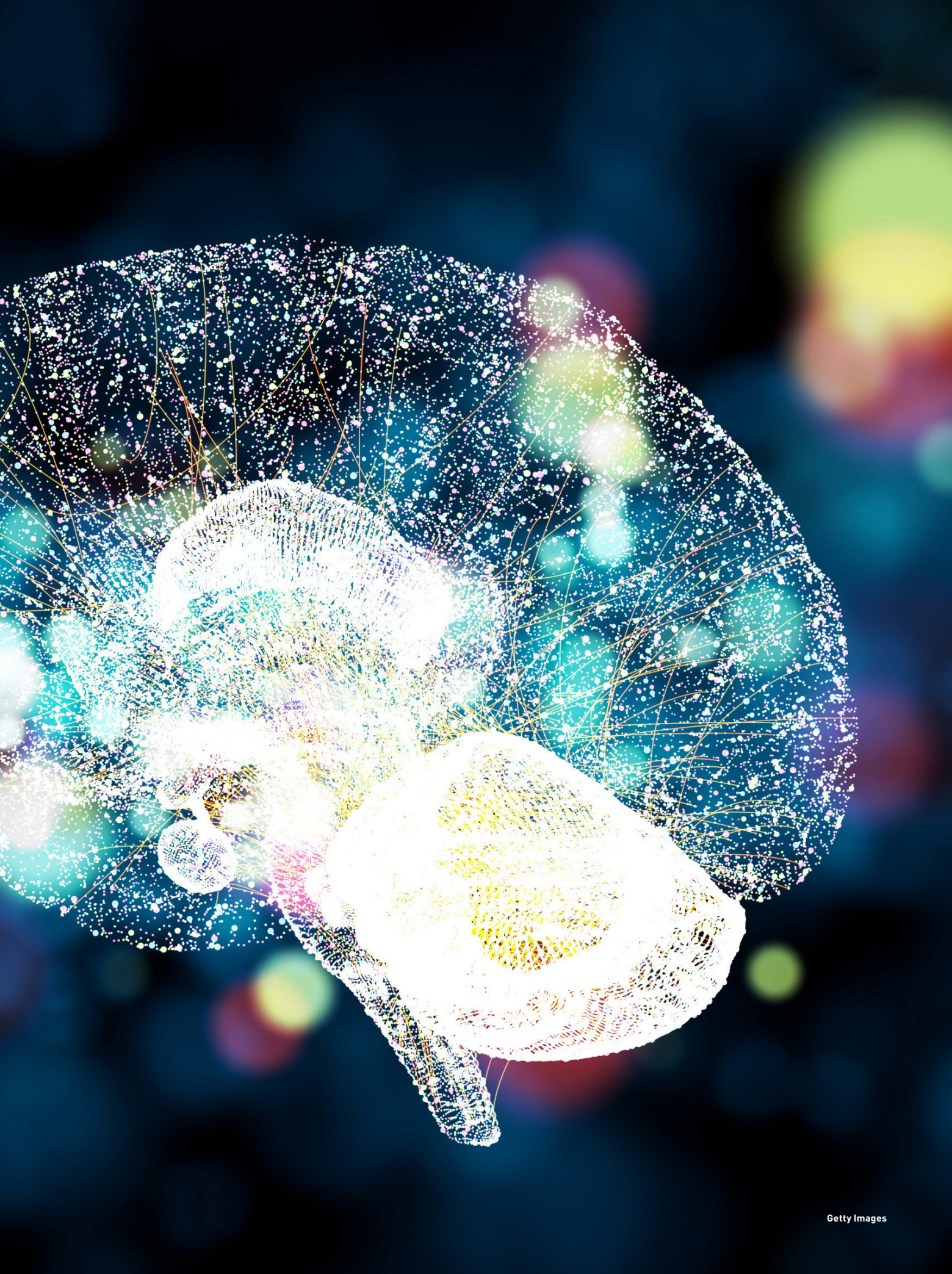
Optical Neural Networks

Edwin Cartlidge

Light-based computers inspired by the human brain could transform machine learning—if they can be scaled up.





Getty Images

Google has developed a neural-network chip to help keep AI applications from overwhelming its data centers—but the chip is still digital.

Google

Recognizing faces in photos, monitoring credit-card transactions for fraud, recommending music based on personal taste and identifying tumors in medical images—artificial intelligence has made huge strides in recent years. These advances and many others have come about largely thanks to progress in neural-network computing and "deep learning."

Somewhat akin to how the human brain works, these networks tune the connections between large numbers of artificial neurons to spot patterns in data sets. Though first developed in the 1950s, artificial neural networks (ANNs) have really taken off only in the last decade or so, according to Jeffrey Shainline of the U.S. National Institute of Standards and Technology (NIST) in Boulder, CO, USA. He attributes that growth to improved training algorithms, more powerful computers and a flood of internet data.

Nevertheless, their many internal connections and need for extensive training can make ANNs very demanding computationally. Several years ago, for example, Google realized that its ever-growing use of AI could lead to a huge increase in energy consumption at its data centers. In response it created a new "Tensor Processing Unit" that saves energy by foregoing universal programmability and focusing instead on the matrix operations that lie at the heart of ANNs.

As powerful as the chip might be, however, it is still based on the basic architecture developed in the 1940s by John von Neumann. That architecture involves wires ferrying data and instructions back and forth between memory and processor—a precise but step-by-step process. In contrast, ANNs, like biological brains, are decentralized and inherently parallel. Thus, when run on digital computers, they have to be simulated using software (even in the case of IBM's TrueNorth chips, which process data within distributed memory units).

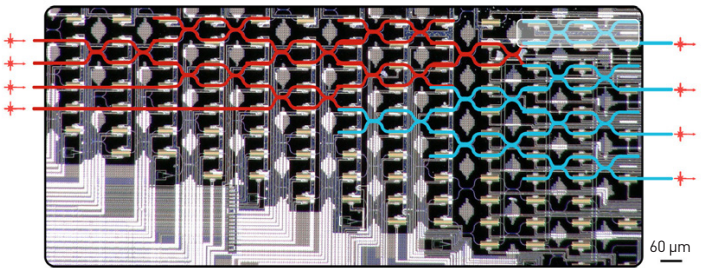
Analog optical technology instead allows ANNs to be implemented directly in hardware, with data encoded in pulses of light and neurons made from beam splitters, waveguides and other components. Freed from the constraints of a clock cycle, argues Junwei Liu, a physicist at the Hong Kong University of Science and Technology, these circuits can potentially be much quicker and more energy-efficient than existing networks. As he explains, their processing time is set not by the number of neural links but simply by how long it takes light to traverse the circuit. "And light," Liu points out, "is very fast."

1047-6938/20/06/32/8-$15.00 © OSA

Optical technology allows ANNs to be implemented directly in hardware, with data encoded in light pulses and neurons made from beam splitters, waveguides and other components.

Coherent solutions

Following the laser's invention in 1960, scientists started to dream of building an optical computer. The emphasis initially was on a digital device, which would use the optical equivalent of a transistor to switch beams of light on and off. Such a machine, claimed its proponents, might operate much faster than an electronic computer, thanks both to the far higher raw speed of light through a circuit and to the possibilities for parallel processing, since photons don't interact with one another.



Researchers at MIT have used a photonic processor containing Mach-Zehnder interferometers to carry out matrix multiplication, consisting of unitary transformations (red) and attenuation (blue).

Reprinted with permission from Y. Shen et al., Nat. Photon. **11**, 441 (2017)

However, this very non-interaction of photons itself created a problem, as it means that they couldn't be used directly to control the behavior of other photons. Instead, switching had to be done indirectly, by modifying some kind of intermediate material. But the power needed to achieve such nonlinearity was so high that many proposed optical schemes were impractical for more than a handful of logic gates. Unable to build an optical transistor that could come close to outperforming its electronic equivalent, many researchers left the field in the 1970s.

Yet these problems didn't spell the end of optical computing. As Alexander Tait at NIST in Boulder explains, optical ANNs have less need for switching than digital computers. In an ANN, each neuron receives multiple weighted (linear) inputs, but generates just one (nonlinear) output (see "The power of learning," p. 36). So whereas a gate-based circuit contains large numbers of nonlinear elements (an AND gate, for example, consists of two transistors in series), the nonlinearity in an ANN is restricted to neurons' output. And that remains true, Tait points out, no matter how many connections there are to a given neuron.

Physicists have devised numerous ways of realizing ANNs optically, one of which uses Mach-Zehnder interferometers (MZIs) to calculate matrix products. By interfering a coherent pair of incoming light pulses having introduced a specific phase shift between them, these devices multiply a two-element vector, encoded in the amplitude of the pulses, by a 2×2 matrix. An array of the interferometers can then perform arbitrary matrix operations.

These operations were first carried out using meter-length bulk optics, but advances in integrated photonics have shrunken things considerably. In 2015, groups at the University of Bristol, U.K., and the Massachusetts Institute of Technology (MIT), USA, announced independently that they had made "nanophotonic processors" capable of carrying out general matrix operations that could potentially be applied to a variety of problems in classical and quantum physics.

The MIT groups, led by Dirk Englund and Marin Soljačić, went on to use 56 MZIs from part of their nanophotonic processor to implement two layers of four neurons. Encoding four different vowel sounds spoken by 90 different people in laser pulses, and using half the data for training, they found that the network identified the sounds correctly 77% of the time, compared with 92% using a conventional 64-bit digital computer.

Despite these positive results, the scheme faces major challenges. For one thing, says Tait, scaling up to larger numbers of MZIs will be difficult, partly because the phase shifters require lots of power. Then there is the question of the nonlinear operation needed to link one set of MZIs with another, which the MIT researchers simply simulated using a normal computer. According to Tait, this nonlinearity would corrupt the light's phase, ruining the calculations. Adherents of this technology, he says, "have yet to propose how the MZIs will cascade from one to another."

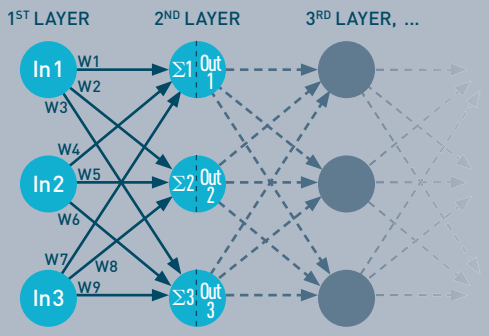
The power of learning

Both natural and artificial neural networks recognize patterns in data by adjusting the strength of connections between small processing units known as neurons. The link between each pair of neurons (known as a synapse in biological brains) is represented by a weight, with every neuron generating an output signal that depends on the sum of its weighted inputs. In the simplest case, the artificial neuron will "fire" when that sum exceeds some specific threshold. More generally, its output will be the result of a nonlinear function whose input is the weighted sum.

Neural networks used for machine learning are commonly arranged in layers: an input layer, several "hidden" layers, and an output layer, with data processed one layer at a time. If every layer contains N neurons, its processing can be represented by two steps. First, it multiplies an N-element input vector by an $N \times N$ weighting matrix, and then passes the result through a nonlinear function to generate another N-element output vector. That output serves as the next layer's input, and the process repeats until the network as a whole generates its N-valued output.

To be trained, a network is typically fed many examples of a certain kind of data set, such as the pixel values of photos of cats, along with a signal at the appropriate output, indicating "cat." With the weights initially assigned random values, the output signal is random. So the weights are adjusted to try and make the output more cat-like the next time around. This process is repeated many times until it reaches a steady state, with every input image of a cat then, in principle, being recognized as such by the network. At that point, the system has learned to distinguish a cat from a dog—or anything else.

Neural network with 3 neurons per layer



Going from one layer to the next is a two-step process:

1. Matrix multiplication

$$(\text{In}\,1 \;\; \text{In}\,2 \;\; \text{In}\,3)\begin{pmatrix} W1 & W2 & W3 \\ W4 & W5 & W6 \\ W7 & W8 & W9 \end{pmatrix} = (\Sigma 1 \;\; \Sigma 2 \;\; \Sigma 3)$$

2. Nonlinear function

$$\left.\begin{aligned} \emptyset(\Sigma 1) &= \text{Out}\,1 \\ \emptyset(\Sigma 2) &= \text{Out}\,2 \\ \emptyset(\Sigma 3) &= \text{Out}\,3 \end{aligned}\right\}$$ (Out 1, Out 2, Out 3) then becomes input for next layer

Opening up the spectrum

A very different approach to all-optical neural networking has been taken by Wolfram Pernice of the University of Münster, Germany, and colleagues. Rather than interfering individual coherent pulses of light, Pernice's team exploits wavelength-division multiplexing (WDM) to transport and sum multiple pulses at different wavelengths using single waveguides. They also use a phase-change material (PCM) for both linear summing and nonlinear firing. Employed in rewritable optical disks, this material transmits far more light when amorphous than when in a crystalline state.

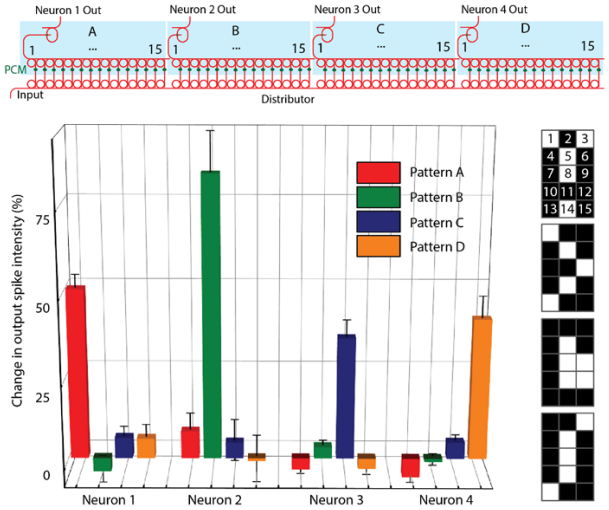
Each neuron in a layer within this scheme uses a series of tiny ring-shaped resonators of varying diameters to tap light signals with corresponding wavelengths travelling together in a common waveguide. The signals at different wavelengths within a given neuron are each attenuated a certain amount by a piece of PCM in their path, and therefore weighted, before being combined in another waveguide. If the total power of all those signals exceeds a certain threshold, they then switch another piece of PCM, this time embedded in a resonator at the neuron's output.

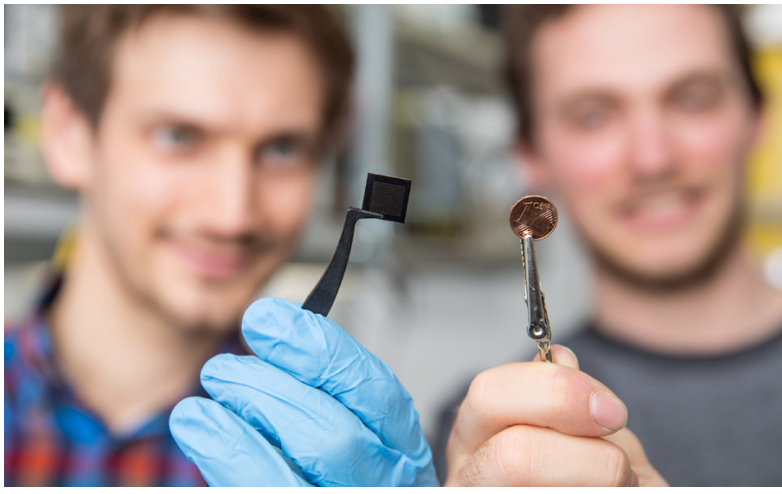
When switched, this piece of PCM passes on a light pulse that would otherwise couple to the resonator, allowing the neuron to fire. The pulses from each of the neurons firing in that layer—each having a different wavelength—are then collected by another waveguide and constitute the input signals to the following layer. This process repeats layer by layer until pulses are generated at the device's output.

Pernice's group showed that it could train a single layer of four neurons, each with 15 inputs, to distinguish between four different pixilated letters, A, B, C and D (see diagram, p. 37). They also showed, using a simpler four-input neuron, how to carry out "unsupervised learning." Using a feedback loop, the neuron reinforced connections to inputs that contributed to a firing, allowing it to spot recurring but unflagged patterns within incoming data.

Pernice says that their system could potentially outpace electronic ANNs and is designed to be scaled up. But he cautions that manufacturing large quantities of very finely-tuned rings (whose resonant wavelengths would need to match to fractions of a nanometer) will not be easy. And Geoffrey Burr, a researcher at IBM in Almaden, California, adds that the neurons' firing rate will in practice be limited by the relatively slow recrystallization time of the PCM.

Problems with all-optical ANNs have led some groups to investigate optoelectronic schemes in which neurons convert signals from light into electricity and then back to light.





Researchers at the University of Münster, Germany, and the University of Oxford and University of Exeter, U.K., developed an optical chip (right) that implements four artificial neurons in photonic waveguides and ring resonators, and that can be taught to distinguish four different letters (left). WWU - Peter Leßmann / Johannes Feldmann

Other groups have proposed different all-optical schemes, each with its own strengths and weaknesses. Liu and several colleagues in Hong Kong have built an ANN with 22 neurons using spatial light modulators and Fourier lenses for matrix multiplication. For nonlinearity, they fix the strength of a laser probe passing through a gas of ultracold atoms using a second laser beam that tunes the relative population of atoms in different energy levels. Liu says they currently aim to realize at least 1000 neurons for solving practical problems such as image recognition, but adds that they need to make their system more efficient to contain costs.

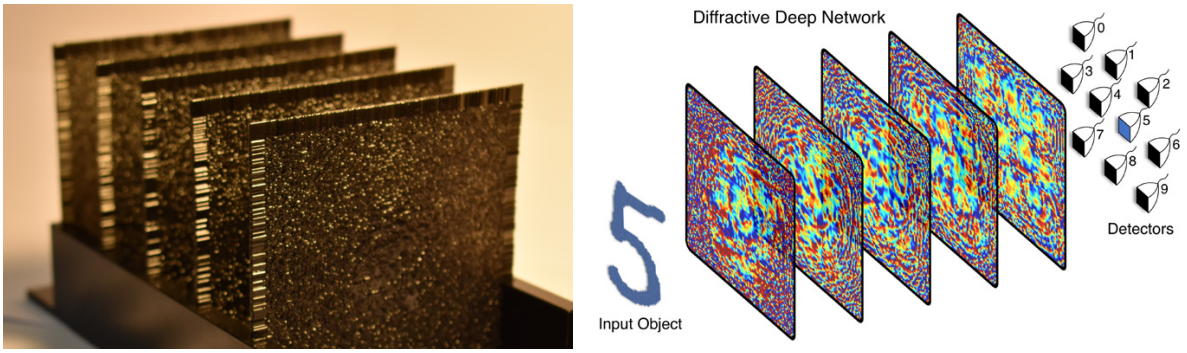
Researchers led by OSA Fellow Aydogan Ozcan at the University of California, Los Angeles, USA, meanwhile, have created hundreds of thousands of neurons from several very thin layers of polymer wafer. Using a computer model and 3D printing, they varied the thickness of the polymer such that it diffracts and directs incoming light to specific points on the final layer. Ozcan is confident that the system, tested out with terahertz radiation, can be modified to work in the visible or infrared, but acknowledges that reprogramming, which currently involves physically altering the layers, is tricky.

Electrons' helping hand

Problems with all-optical ANNs, in particular how to implement nonlinearity, have led some groups to investigate optoelectronic schemes in which neurons convert signals from light into electricity and then back to light—the electrical signals being amenable to nonlinear operations. Scientists at Princeton University, USA, for example, exploit WDM to connect multiple neurons with a single waveguide, while using an electro-optic modulator to introduce nonlinearity.

The Princeton scheme sets weights by heating micro rings, which former group member Bhavin Shastri, now at Queen's University in Kingston, Canada, says allows the weights to be changed very quickly (although they do require a continuous supply of energy). As such, he reckons the approach might enable learning on the fly, conceivably in the fast-changing environments of self-driving cars.

Another potential application is in compensating nonlinear distortions of signals in long-distance fiber optic cables. This can be done with conventional electronics, either deterministically or using neural networks, but high-speed transmission entails significant power consumption. The Princeton group, working



UCLA researchers identified handwritten digits by illuminating them and diffracting the light through a series of thin polymer wafers to trigger one of ten detectors. Ozcan Research Group/UCLA

with researchers at NEC Corp. in the United States and Japan, recently showed that a WDM-based photonic neural network could create an effective model of nonlinear distortions in a 10,000-kilometer-long stretch of trans-Pacific fiber accurately and efficiently.

Optoelectronic technology can also be used to implement a very different type of neural network known as a "reservoir computer", in which most weights are never adjusted at all. This kind of network, which has significant practical potential, uses a random matrix to look for features in time-dependent data (see "A random reservoir," p. 39).

The French company LightOn also exploits random matrices, but does so with what it calls an optical processing unit. This uses a camera to analyze laser light that has been encoded with image data via a micromirror array and that's then passed through a diffusive medium. The scattered light interferes randomly on the camera to generate a speckle pattern, with the brightness of the pixels (a nonlinear function of the speckle field) used to identify specific features in the images. The company's chief technology officer, Laurent Daudet, says that the processing unit—due to be made widely available online in April—is designed to complement digital processors, allowing quicker, more efficient matrix calculations.

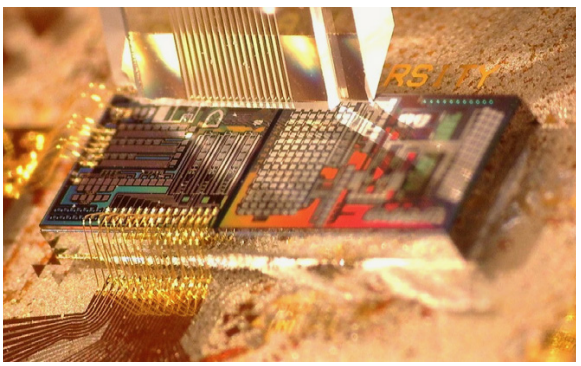
Thinking big

Despite the enthusiasm of many researchers toward optical ANNs, some scientists are skeptical that they can compete with digital computers. Rodney Tucker, an electrical engineer at the University of Melbourne, Australia, argues that electronic networks are easier to program and scale up. He also maintains that they are in fact more energy efficient, given the losses in optical systems—particularly for large-scale computations. "It is the energy per operation that matters," he says.

Even those working on optical ANNs admit that significant hurdles remain. Shastri says that two things must happen before optoelectronic devices can compete with electronic ones. One is creating a scalable platform, such as a silicon substrate, that can support both photonic and electronic components. The other is the ability to integrate lasers onto such a platform cheaply. He notes that silicon lases only at very low temperatures, while indium phosphide lasers are more expensive.

For Charles Roques-Carmes at MIT, such hurdles need to be overcome in the next few years; otherwise rich multinationals will probably find a way to further improve electronic devices. "If optics is to make a difference it has to make it quite fast," he says.

Nevertheless, with Moore's law under threat as electronic circuits reach the limits of miniaturization, optical technologies have attracted significant interest and some investment, even by apparent competitors. The researchers at MIT have spun out two companies to develop chips that use optics to boost the speed of matrix multiplication. Lightelligence has raised at least US$10 million in funding, while Lightmatter has received



Micrograph image of a wirebonded photonic neural network chip from Princeton University's Lightwave Lab.
C. Huang et al., OFC 2020 / Princeton University Lightwave Lab

With Moore's law under threat as electronic circuits reach the limits of miniaturization, optical technologies have attracted significant interest and some investment.

more than US$30 million—including money from GV, a venture arm of Google's parent company Alphabet.

The research at Princeton has also led to the creation of a new company, Luminous Computing, which has attracted at least US$9 million in seed funding. Tait, who did his Ph.D. with the Princeton group, thinks this is a good sign for the field. "The start-ups being founded is important because there are now big teams of engineers working on these problems," he says.

But not all groups are thinking about commercial products. Tait is working on a NIST project headed by Shainline that is developing hardware designed, says Shainline, to "be scalable up to the human brain or beyond." That entails potentially building networks with tens of billions of neurons, which, Shainline says, means the neurons must be as efficient as possible to limit energy consumption (the human brain requiring a mere 20 W to operate). As such, the group has imposed a strict design criterion: communication between neurons must occur at the level of single photons.

The scheme on the drawing board relies on superconductors to detect photons and to update weights and sum inputs; semiconductors to generate the photons; and photonic components to distribute them—all at just a few degrees above absolute zero. Shainline acknowledges that combining these different components will not be easy, but says that he and his four colleagues will take their time to get the technology right.

He estimates they will have a single neuron ready in a couple of years' time, a functioning network with several hundred neurons on a single chip within about five years, and, if all goes well, multi-wafer systems comprising billions of neurons after a decade or more. At that point, he enthuses, they should have an "extraordinarily powerful" machine. "In the biological brain the neurons and synapses have rich complexity and most hardware projects miss out on that," he says. "But I think that is what makes the brain work so well." OPN

Edwin Cartlidge (edwin.cartlidge@yahoo.com) is a freelance science journalist based in Rome.

For references and resources, go online:
www.osa-opn.org/link/neural-nets

A random reservoir

To analyze data that are changing in time, a neural network can be made "recurrent" by connecting neurons not simply layer by layer, but to themselves and others some distance away. This lets signals travel backwards as well as forwards through the network, setting up feedback loops and the potential for a network "memory"—but also vastly complicating the nonlinear-weighting problem of training the network.

One way around this problem is a so-called reservoir computer. Such a computer consists of a recurrent network with randomized weights, as well as one additional layer of neurons at the input and another at the output. The idea is that, given a big enough network, somewhere within it there will be a region of connected neurons with the right weights to recognize signals of interest. Training involves a simple linear matrix multiplication to set the weights of the output neurons, while the weights within the recurrent network itself are left alone.

According to Guy Van der Sande of the Vrije Universiteit Brussel in Belgium, reservoir networks could prove particularly handy in signal analysis for optical routing, as they could eliminate the need to convert signals from light into electricity and then back to light, thereby saving time and energy. Another use could be reconstructing distorted signals from optical fibers or overcrowded cellular networks.

Toward another application, the team of Damien Rontani, University of Lorraine, France, has built a reservoir computer that can successfully distinguish between six human actions: walking, running, jogging, waving, clapping and boxing. The optoelectronic system uses a standard computer to assign values of pixels in a spatial light modulator on the basis of video footage of the six actions, plus camera to record light from an LED that is reflected off the modulator. The recorded images constitute the processed data, while the pixels represent the neurons – their values given a random component by the computer.