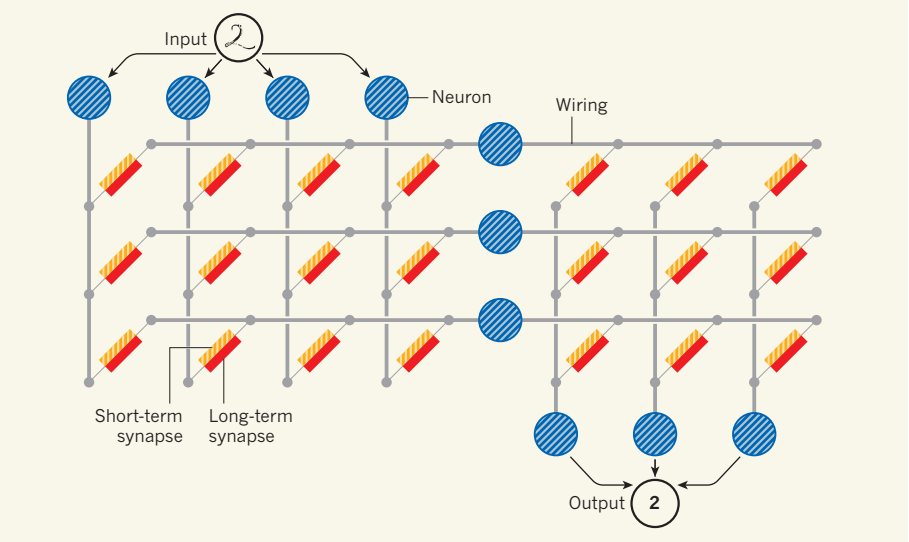

Figure 1 | An artificial neural network containing two types of synapse. Ambrogio *et al.*[1] report a hardware-based artificial neural network that is trained to classify complex images, such as handwritten numbers, with an accuracy similar to that of a software-based network. The network consists of artificial neurons linked by wires to two types of artificial synapse (connections between neurons). Short-term synapses (which can retain alterations in their synaptic state for milliseconds) are used regularly during network training, whereas long-term synapses (with state retention of years) are used mainly for memory. The long-term synapses are physical devices, whereas the neurons and short-term synapses are simulated computationally (indicated by hatching).

as phase-change-memory (PCM) devices, show promise because of their small physical size and excellent retention properties. PCM devices contain a material known as a chalcogenide glass, which can switch reversibly between an amorphous phase (of high resistance) and a crystalline phase (of low resistance). The device's resistance state is programmed by crystallizing part of the material using local heating produced by an applied voltage. This state is retained long after the voltage has been removed, and further programming can be achieved by crystallizing other parts of the material.

Unfortunately, PCM devices can be programmed in only one direction: from high to low resistance, by changing from low to high crystallinity. To achieve the desired resistance state with good precision, sequences of hundreds of voltage pulses are required. If the desired state is overshot, the chalcogenide glass must be completely reset to the amorphous phase and the step-by-step programming restarted. This shortcoming, combined with variations between devices caused by the manufacturing process, can slow or even prevent network training, as previous work by the group that performed the current study has shown[2]. As a result, the prototype networks that have been constructed using these devices[3,4] are impractical and have much lower image-classification accuracies than do software-based networks.

The breakthrough of Ambrogio and colleagues' work lies in a two-tier, bio-inspired approach. In biological neural networks, short-term changes in the states of synapses support a variety of computations, whereas long-term changes provide a platform for learning and memory[5]. For this reason, the authors' artificial neural network uses synaptic 'cells' that contain two types of synapse: short-term and long-term (Fig. 1).

The short-term synapses are used regularly during network training. They require only brief state retention, but fast and precise programming to the desired state. Such features are provided by an electronic switch called a transistor, which has a capacitor (a device for storing electric charge) attached to one of its electrodes, known as the gate[6]. The transistor's state is programmed by a fast voltage pulse applied to the gate. The capacitor maintains this voltage for a few milliseconds, providing brief state retention.

After the network has been trained on several thousand images and the short-term synapses have changed states substantially, the synaptic states are written into long-term synapses. The cycle is then repeated until all of the training images have been presented to the network. The long-term synapses are used for network operation after training is complete. They consist of PCM devices that have state-retention times of years, at the expense of tedious, energy-intensive programming.

An advantage of this technique is that the transfer of states from short- to long-term synapses can be done in electronic-circuit blocks separate from the network, while the network carries out other tasks. Moreover, although the authors' synaptic cells are more complicated in practice — containing one capacitor, two PCM devices and five transistors — they are still about half the size of artificial synapses used in other networks[6].

Ambrogio *et al.* tested their synaptic-cell approach using a fairly complex artificial neural network containing multiple layers of neurons and more than 200,000 PCM devices. The authors carried out classification tasks using three standard sets of images: greyscale handwritten numbers from the MNIST database[7], and colour images from the CIFAR-10 and CIFAR-100 databases[8]. The accuracies obtained were 98%, 88% and 68%, respectively. These results are strikingly similar to those obtained using TensorFlow, a leading neural-network software (see www.tensorflow.org).

Despite these impressive findings, a key limitation of the work is that only the PCM devices were actually fabricated; the other components of the synaptic cells and the neurons were simulated computationally. The authors took care to use accurate models that consider variations between transistors, and they proposed a method to minimize the impact of such variability on synaptic-cell performance. Most importantly, they carried out a detailed power assessment, and found that their proposed technology would consume about 100 times less power than current state-of-the-art networks, while providing a similar classification performance. Nevertheless, only a working hardware prototype will convince industry of the technology's performance and low-power advantages. Furthermore, the estimated power consumption is still a far cry from that of biological neural networks, leaving plenty of room for improvement.

However, Ambrogio and colleagues' work is more than a crucial stepping stone to the integration of PCM devices in neural-network hardware. It will also inspire device research, because it creates a need for nanoscale short-term synapses to replace the bulky transistor–capacitor ones. A wall in emerging memory technologies has been breached — networks based on these devices can work as well as do their software counterparts. This finding suggests that advances in artificial intelligence will not only continue, but also be accelerated by emerging hardware. ■

Gina C. Adam *is at the National Institute for Research and Development in Microtechnologies, Bucharest 077190, Romania.*
e-mail: gina.adam@imt.ro

1. Ambrogio, S. *et al. Nature* **558**, 60–67 (2018).
2. Burr, G. W. *et al. 2014 IEEE Int. Electron Devices Meet.* 29.5.1–29.5.4 (IEEE, 2014).
3. Gokmen, T. & Vlasov, Y. *Front. Neurosci.* **10**, 333 (2016).
4. Yu. S. *et al. 2015 IEEE Int. Electron Devices Meet.* 17.3.1–17.3.4 (IEEE, 2015).
5. Abbott, L. F. & Regehr, W. G. *Nature* **431**, 796–803 (2004).
6. Kim. S., Gokmen, T., Lee, H.-M. & Haensch, W. E. *2017 IEEE 60th Int. Midwest Symp. Circuits Systems* 422–425 (IEEE, 2017).
7. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. *Proc. IEEE* **86**, 2278–2324 (1998).
8. Krizhevsky, A. *Learning Multiple Layers of Features From Tiny Images* Ch. 3; www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf; https://www.cs.toronto.edu/~kriz/cifar.html (2009).