doi:10.1038/nature14441

Training and operation of an integrated neuromorphic network based on metal-oxide memristors

M. Prezioso[1]*, F. Merrikh-Bayat[1]*, B. D. Hoskins[1]*, G. C. Adam[1], K. K. Likharev[2] & D. B. Strukov[1]

Despite much progress in semiconductor integrated circuit technology, the extreme complexity of the human cerebral cortex[1], with its approximately 10^{14} synapses, makes the hardware implementation of neuromorphic networks with a comparable number of devices exceptionally challenging. To provide comparable complexity while operating much faster and with manageable power dissipation, networks[2] based on circuits[3,4] combining complementary metal-oxide-semiconductors (CMOSs) and adjustable two-terminal resistive devices (memristors) have been developed. In such circuits, the usual CMOS stack is augmented with one[3] or several[4] crossbar layers, with memristors at each crosspoint. There have recently been notable improvements in the fabrication of such memristive crossbars and their integration with CMOS circuits[5–12], including first demonstrations[5,6,12] of their vertical integration. Separately, discrete memristors have been used as artificial synapses in neuromorphic networks[13–18]. Very recently, such experiments have been extended[19] to crossbar arrays of phase-change memristive devices. The adjustment of such devices, however, requires an additional transistor at each crosspoint, and hence these devices are much harder to scale than metal-oxide memristors[11,20,21], whose nonlinear current–voltage curves enable transistor-free operation. Here we report the experimental implementation of transistor-free metal-oxide memristor crossbars, with device variability sufficiently low to allow operation of integrated neural networks, in a simple network: a single-layer perceptron (an algorithm for linear classification). The network can be taught *in situ* using a coarse-grain variety of the delta rule algorithm[22] to perform the perfect classification of 3 × 3-pixel black/white images into three classes (representing letters). This demonstration is an important step towards much larger and more complex memristive neuromorphic networks.

In a hybrid CMOS/memristor circuit, the CMOS subsystem contacts each wire, and hence can address each memristor on the add-on crossbar(s), using a specific 'CMOL' area-distributed interface[3,4]. The basic idea of hybrid neuromorphic networks—so-called CrossNets[2]—is to use this opportunity to connect CMOS-implemented hardware models of neuron bodies with the memristive crossbar(s), whose wires play the parts of axons and dendrites and whose memristors mimic biological synapses. The simple, two-terminal, transistor-free topology of metal-oxide memristors may enable CrossNets to achieve extremely high density—much higher than that of pure-CMOS neuromorphic networks (including those based on CMOS-modelled memristors[23], floating-gate[24] and ferroelectric[25] memory cells), and even higher than that of their biological prototypes. For example, a CrossNet based on a hybrid CMOS/memristor circuit with five layers of 30-nm-pitch crossbars, two memristors per synapse, and 10^4 synapses per neural cell would have an areal density of about 25 million cells per square centimetre, that is, higher than that in the human cerebral cortex, at comparable average connectivity[1]. Estimates show that such CrossNets may also provide comparable power efficiency, at a much higher operation speed—for example, an intercell signal transfer delay of about 0.02 ms (compared to about 10 ms in biological systems) at an easily manageable energy dissipation rate of about 1 W cm^{-2}.

However, the practical implementation of such networks is still very challenging, owing to the specific physical mechanism of resistance change in most prospective metal-oxide-based memristors—a reversible modulation of the concentration profile of oxygen vacancies[11,20,21]. On the positive side, the atomic scale of the vacancy position modulation implies the possibility of memristor scaling down to few-nanometre dimensions, which has been confirmed by recent experiments[26,27]. On the negative side, such a small scale makes the device-to-device reproducibility of device parameters, most importantly the voltage required for memristor electroforming and switching[20,21], difficult to achieve with the currently used fabrication technologies. Device variability is the main reason why the only (to our knowledge) demonstrations of memristive neuromorphic networks were based on disconnecting each memristor from the crossbar for individual forming, using either a crossbar with external (off-chip) wires[18], or an individual switch transistor at each crosspoint[19]. Both these approaches are incompatible with the goal of reaching the extremely high density of neuromorphic networks discussed above.

The main goal of this work was an experimental demonstration of a fully operational neural network based on an integrated, transistor-free crossbar with metal-oxide memristors. To reach this goal, a large reduction of memristor variability was essential, and to achieve it, we used binary-oxide $\mathrm{Al_2O_3/TiO_{2-x}}$ stacks (see inset to Fig. 1b). Their fabrication procedure was generally close to that described in ref. 27, but with the important difference of using low-temperature (<300 °C) reactive sputtering for film deposition, which enables monolithic three-dimensional integration. The stack was first optimized by conducting an exhaustive experimental search over a range of titanium dioxide compositions and layer thicknesses (from 5 nm to 100 nm) to find the parameter range providing the lowest forming voltages. Within that range, the device performance—most importantly the memristor uniformity and the current–voltage curve nonlinearity—was further optimized by varying the aluminium oxide thickness from 1 nm to 5 nm (Supplementary Information Section 1).

The main feature of such optimized junctions is their low variability (Supplementary Figs 3 and 4). In addition, other important characteristics of the 200 nm × 200 nm formed devices are also desirable: the ON/OFF current ratios of above four orders of magnitude (at ±0.1 V), high nonlinearity of the current–voltage curves (with the current at the switching voltage more than ten times the current at half of the switching voltage), a switching endurance of at least 5,000 cycles, an estimated memory retention of at least ten years at room temperature, low forming (~2 V) and switching (~1.5 V) voltages, and relatively low operation currents of between ~100 nA and ~100 μA (see Supplementary Fig. 1).

The optimized technology was then used to fabricate an integrated memristive crossbar with 12 × 12 devices (Fig. 1), with a few process

[1] Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, California 93106, USA. [2] Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, USA.

*These authors contributed equally to this work.

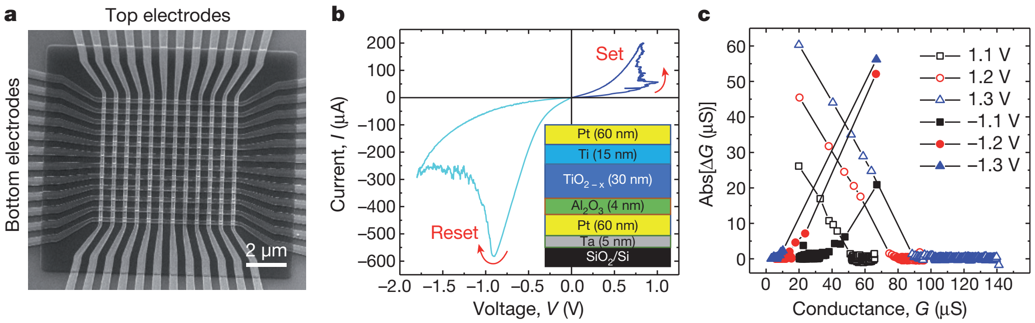

Figure 1 | Memristor crossbar. a, Integrated 12 × 12 crossbar with an Al_2O_3/TiO_{2-x} memristor at each crosspoint. **b**, A typical current–voltage curve of a formed memristor. **c**, Absolute values of conductance change under the effect of 500-μs voltage pulses of two polarities, as a function of the initial conductance, for various pulse amplitudes. The inset in **b** shows the device cross-section schematically.

modifications to increase the metal electrode thickness, so that the line resistances were reduced to about 800 Ω for the top layer of the crossbar and 600 Ω for its bottom layer. The crossbars retained the excellent uniformity of virgin (pre-formed) crossbar-integrated devices (see Supplementary Figs 3, 4 and 5), allowing individual electric forming and tuning of each memristor. The electroforming was performed by grounding the corresponding bottom electrode and applying a current-controlled ramp-up to the top electrode, while leaving all other line potentials floating (Supplementary Fig. 4). To minimize current leakage during the subsequent forming of other devices, each formed memristor was immediately switched into its low-current (OFF) state. The measured individual characteristics of the formed memristors were mostly similar to those of stand-alone devices, except for a somewhat smaller (∼100) ON/OFF current ratio. This difference may be partly explained by current leakage through other crosspoints at the measurements, and partly by the somewhat smaller switching voltages used for the crossbar to lower the risk of device damage. In addition, some deviations from the optimal device performance could be caused by the electron-beam evaporation of thicker electrodes, which required breaking of the vacuum, as opposed to the fully *in situ* sputtering of single device layers, and their subsequent annealing (see Supplementary Information).

The fabricated memristive crossbar was used to implement a simple artificial neural network with the top-level (functional) scheme shown in Fig. 2. This is a single-layer perceptron[22] with ten inputs and three outputs, fully connected with 10 × 3 = 30 synaptic weights (Fig. 2b). As the scheme shows, the perceptron's outputs f_i (with $i = 1, 2, 3$) are calculated as nonlinear 'activation' functions:

$$f_i = \tanh(\beta I_i) \tag{1}$$

of the vector-by-matrix product components:

$$I_i = \sum_{j=1}^{10} W_{ij} V_j \tag{2}$$

Here V_j with $j = 1,\ldots,9$ are the input signals, V_{10} is a constant bias, β is a parameter controlling the function's nonlinearity, and W_{ij} are adjustable (trainable) synaptic weights. Such a network is sufficient for performing, for example, the classification of 3 × 3-pixel black-and-white images into three classes, with nine network inputs ($V_1,\ldots,V_9$) corresponding to the pixel values. We tested the network on a set of $N = 30$ patterns, including three stylized letters ('z', 'v' and 'n') and three sets of nine noisy versions of each letter, formed by flipping one of the pixels of the original image (see Fig. 2c). Because of the very limited size of the set, it was used for both training and testing.

Physically, each input signal was represented by a voltage V_j equal to either +0.1 V or −0.1 V, corresponding, respectively, to the black or white pixel, while the bias input V_{10} was equal to −0.1 V. Such coding makes the benchmark input set balanced, in particular ensuring that the sum of all input signals across all patterns of a particular class is close to zero, which speeds up the convergence process[28]. To sustain this balance at the network's output as well, each synapse

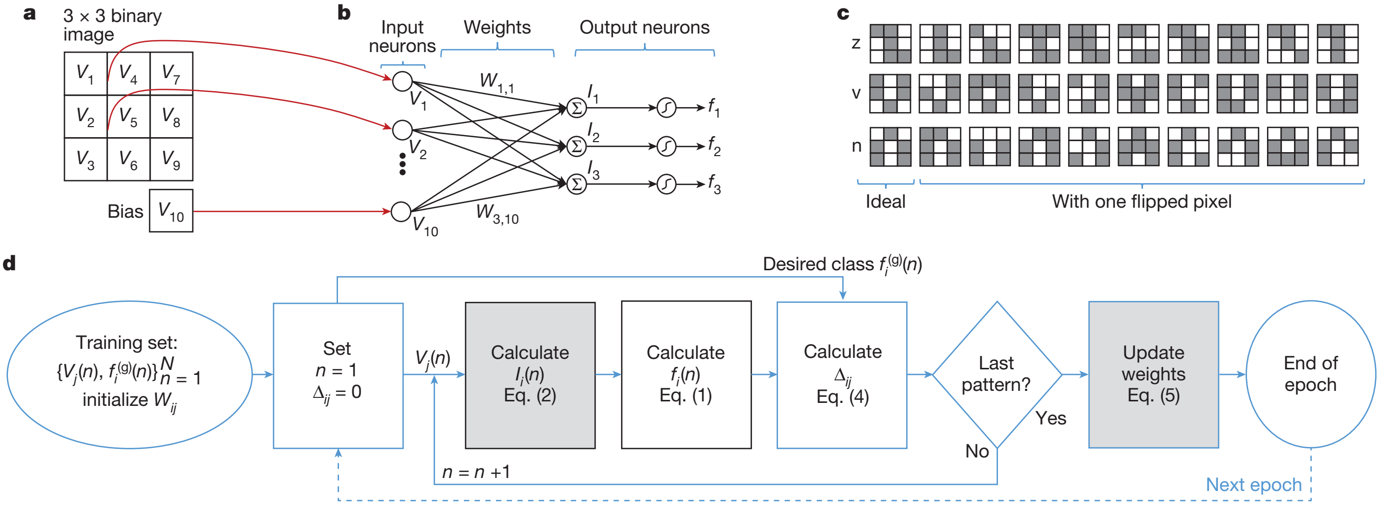

Figure 2 | Pattern classification experiment (top-level description). a, Input image. **b**, The single-layer perceptron for classification of 3 × 3 binary images. **c**, The used input pattern set. **d**, The flow chart of one epoch of the used *in situ* training algorithm. In **d**, the grey-shaded boxes show the steps implemented inside the crossbar, while those with solid black borders denote the only steps required to perform the classification operation.

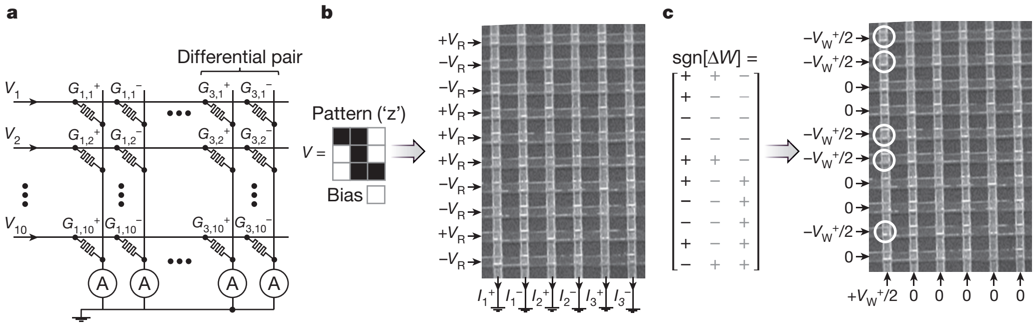
a
Differential pair
V_1
V_2
V_{10}
$G_{1,1}^+$ $G_{1,1}^-$ $G_{3,1}^+$ $G_{3,1}^-$
$G_{1,2}^+$ $G_{1,2}^-$ $G_{3,2}^+$ $G_{3,2}^-$
$G_{1,10}^+$ $G_{1,10}^-$ $G_{3,10}^+$ $G_{3,10}^-$
A
b
Pattern ('z')
V =
Bias
$+V_R$ $-V_R$ $-V_R$ $+V_R$ $+V_R$ $+V_R$ $-V_R$ $-V_R$ $+V_R$ $-V_R$
I_1^+ I_1^- I_2^+ I_2^- I_3^+ I_3^-
c
sgn[ΔW] =
$-V_W^+/2$ $-V_W^+/2$ 0 0 $-V_W^+/2$ $-V_W^+/2$ 0 0 $-V_W^+/2$ 0
$+V_W^+/2$ 0 0 0 0 0

Figure 3 | Pattern classification experiment (physical-level description). **a**, An implementation of a single-layer perceptron using a 10 × 6 fragment of the memristive crossbar. **b**, An example of the classification operation for a specific input pattern (stylized letter 'z'), with the crossbar input signals equal to $+V_R$ or $-V_R$, depending on the pixel colour. (The read and write biases were always $V_R = 0.1$ V and $V_W^{\pm} = \pm1.3$ V, respectively.) **c**, An example of the weight adjustment in a specific (first positive) column, for a specific error matrix. At the step shown, only the synapses whose weights should be increased (marked by '+' in the table on the left) are adjusted, that is, the memristor conductances $G_{1,1}^+$, $G_{1,2}^+$, $G_{1,5}^+$, $G_{1,6}^+$ and $G_{1,9}^+$ are being increased.

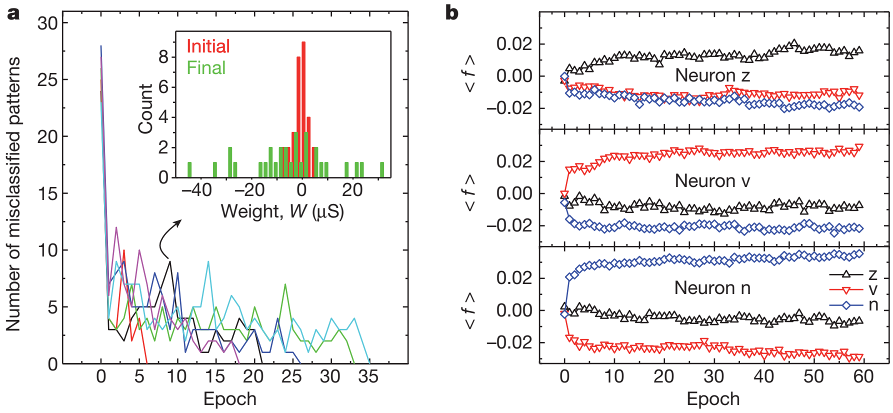

Figure 4 | Pattern classification experiment: results. **a**, Convergence of network outputs, during the training process, to the perfect value (zero), for six training runs from different initial states. **b**, The evolution of output signals, averaged over all patterns of a specific class. The inset in **a** shows the distribution of weights W in the initial state and immediately after epoch 21, when perfect classification is achieved for the first time for this particular run. The classification was considered successful when the output signal f_i corresponding to the correct class of the applied pattern was larger than all other outputs. Such perfect classification was achieved, on average, after 23 epochs, with the standard deviation of ten epochs. The training illustrated by **b** was continued even after the perfect classification had been achieved on epoch 21, to verify that the difference between the output signals continued to increase (unlike the 'perceptron rule' training used in ref. 18).

was implemented with two memristors, so that the total number of memristors in the crossbar was 30 × 2 = 60. Using external electronics to enforce the virtual ground conditions on each column line, and to subtract currents flowing in the adjacent columns to form a differential output signal I_i, we ensured that Ohm's law applied to each column of the crossbar gave a result identical to equation (2), with differential weights:

$$W_{ij} = G_{ij}^+ - G_{ij}^- \qquad (3)$$

where $G_{ij}^{\pm}$ is the effective conductance of each memristor, namely the I/V ratio at voltage 0.1 V. For our devices, these effective conductances were in the range 10–100 μS, so that currents I_i were of the order of a few microamperes. Activation functions—see equation (1)—were also implemented, using external electronics, with the slope $\beta = 2 \times 10^5$ A^{-1} chosen according to the recommendation in ref. 28, confirmed by our own computer simulations (Supplementary Fig. 10).

The network was trained *in situ*, that is, without using its external computer model, using the Manhattan update rule[29], which is essentially a coarse-grain, batch-mode variation of the usual delta rule of supervised training[22]. At each iteration ('epoch') of this procedure, sketched in Fig. 2d, patterns from the training set were applied, one by one, to the network's input, and its outputs $f_i(n)$, where n is the pattern number, were used to calculate the delta-rule weight increments:

$$\Delta_{ij}(n) = \delta_i(n) V_j(n)$$

with

$$\delta_i(n) = \left[f_i^{(g)}(n) - f_i(n)\right] \left.\frac{\mathrm{d}f}{\mathrm{d}I}\right|_{I=I_i(n)} \qquad (4)$$

Here $f_i^{(g)}(n)$ is the target value of the ith output for the nth input pattern. (In our system these values were chosen to be +0.85 for the output corresponding to the correct pattern class, and −0.85 for the output corresponding to the wrong class.) Once all N patterns of the training set had been applied, and all $\Delta_{ij}(n)$ calculated, the synaptic weights were modified using the following Manhattan update rule:

$$\Delta W_{ij} = \eta \,\mathrm{sgn} \sum_{n=1}^{N} \Delta_{ij}(n) \qquad (5)$$

where η is a constant that scales the training rate. (The only difference between the Manhattan update rule from the batch-mode delta rule is the binary quantization, expressed in equation (5) by the 'sgn' function, which simplifies the hardware implementation of the delta rule.

Physically, in our system the weights were modified in parallel for each column of the crossbar (corresponding to a certain value of index i in the above formulas), using two sequential voltage pulses. Namely, first a 'set' pulse with amplitude $V_W^+ = 1.3$ V was applied to increase the conductances of the synapses whose ΔG values, calculated from equation (5), were positive; then a 'reset' pulse $V_W^- = -1.3$ V was applied to the remaining synapses of that column (see Fig. 3c). This fixed-amplitude pulse procedure followed the Manhattan update rule only approximately, because the actual training rate ΔG depends on the initial conductance G of the memristor (see Fig. 1c and Supplementary Fig. 6). (For $G = 20$ μS, ΔG was close to $+60$ μS for the set pulse and -5 μS for the reset pulse, while for $G = 65$ μS, the changes were close, respectively, to $+24$ μS and -55 μS.) Owing to the specific (though quite representative[11]) switching dynamics of our devices, the best classification performance was achieved when the memristors had been initialized somewhere in the middle of their conductance range, around 35 μS (Supplementary Fig. 7b). At such initialization, the perfect classification was reached, on average, after 23 training epochs (see Fig. 4).

In summary, here we have experimentally demonstrated an artificial neural network using memristors integrated into a dense, transistor-free crossbar circuit. This crossbar performed, on the physical (Ohm's law) level, the analogue vector-by-matrix multiplication of equations (2) and (3), which is by far the most computationally intensive part of the operation of any neuromorphic network used repeatedly in the same environment. The other operations, described by equations (1), (4) and (5), were performed by external electronics, but they are much less critical for network performance, and in future, larger CrossNets may be (at least partly) assisted by CMOS subsystems. This is an important step towards the effective analogue-hardware implementation of much more complex neuromorphic networks, from multilayer-perceptron classifiers with deep learning[30] to elaborate CrossNet-based cognitive systems. Recent experiments[27] with similar but smaller (discrete) devices imply that such circuits may be scaled down to devices of 30 nm across or less, that is, to networks with a density of approximately 10^{10} synapses per square centimetre in each crossbar layer.

Received 16 December 2014; accepted 19 March 2015.

1. Mountcastle, V. B. *The Cerebral Cortex* (Harvard Univ. Press, 1998).
2. Likharev, K. K. CrossNets: neuromorphic hybrid CMOS/nanoelectronic networks. *Sci. Adv. Mater.* **3,** 322–331 (2011).
3. Likharev, K. K. Hybrid CMOS/nanoelectronic circuits: opportunities and challenges. *J. Nanoelectron. Optoelectron.* **3,** 203–230 (2008).
4. Strukov, D. B. & Williams, R. S. Four-dimensional address topology for circuits with stacked multilayer crossbar arrays. *Proc. Natl Acad. Sci. USA* **106,** 20155–20158 (2009).
5. Xia, Q. *et al.* Memristor-CMOS hybrid integrated circuits for configurable logic. *Nano Lett.* **9,** 3640–3645 (2009).
6. Chevallier, C. J. *et al.* 0.13μm 64Mb multi-layered conductive metal-oxide memory. *Int. Solid-State Circuits Conf.* **10,** 260–261 (2010).
7. Miyamura, M. *et al.* Programmable cell array using rewritable solid-electrolyte switch integrated in 90 nm CMOS. *Int. Solid-State Circuits Conf.* **11,** 228–229 (2011).
8. Kawahara, A. *et al.* An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput. *Int. Solid-State Circuits Conf.* **12,** 432–434 (2012).
9. Kim, G. H. *et al.* 32×32 crossbar array resistive memory composed of a stacked Schottky diode and unipolar resistive memory. *Adv. Funct. Mater.* **23,** 1440–1449 (2013).
10. Kim, K.-H. *et al.* A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications. *Nano Lett.* **12,** 389–395 (2012).
11. Yang, J. J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. *Nature Nanotechnol.* **8,** 13–24 (2013).
12. Liu, T. *et al.* A 130.7-mm 2-layer 32-Gb ReRAM memory device in 24-nm technology. *IEEE J. Solid-State Circuits* **49,** 140–153 (2014).
13. Jo, S. H. *et al.* Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **10,** 1297–1301 (2010).
14. Chanthbouala, A. *et al.* A ferroelectric memristor. *Nature Mater.* **11,** 860–864 (2012).
15. Seo, K. *et al.* Analog memory and spike-timing-dependent plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device. *Nanotechnology* **22,** 254023 (2011).
16. Ohno, T. *et al.* Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nature Mater.* **10,** 591–595 (2011).
17. Ziegler, M. *et al.* An electronic version of Pavlov's dog. *Adv. Funct. Mater.* **22,** 2744–2749 (2012).
18. Alibart, F., Zamanidoost, E. & Strukov, D. B. Pattern classification by memristive crossbar circuits with ex-situ and in-situ training. *Nature Commun.* **4,** 2072 (2013).
19. Eryilmaz, S. B. *et al.* Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array. *Front. Neurosci.* **8,** 205 (2014).
20. Waser, R., Dittman, R., Staikov, G. & Szot, K. Redox-based resistive switching memories. *Adv. Mater.* **21,** 2632–2663 (2009).
21. Wong, H. S. P. *et al.* Metal–oxide RRAM. *Proc. IEEE* **100,** 1951–1970 (2012).
22. Hertz, J., Krogh, A. & Palmer, R. G. *Introduction to the Theory of Neural Computation* (Perseus, 1991).
23. Pershin, Y. V. & Di Ventra, M. Experimental demonstration of associative memory with memristive neural network. *Neural Netw.* **23,** 881–886 (2010).
24. Hasler, J. & Marr, B. Finding a roadmap to achieve large neuromorphic hardware systems. *Front. Neurosci.* **7,** 118 (2013).
25. Kaneko, Y., Nishitani, Y. & Ueda, M. Ferroelectric artificial synapses for recognition of a multishaded image. *IEEE Trans. Electron. Dev.* **61,** 2827–2833 (2014).
26. Pi, S., Lin, P. & Xia, Q. Cross point arrays of 8 nm × 8 nm memristive devices fabricated with nanoimprint lithography. *J. Vacuum Sci. Technol. B* **31,** 06FA02–1 (2013).
27. Govoreanu, B. *et al.* Vacancy-modulated conductive oxide resistive RAM (VMCO-RRAM). *IEDM Tech Dig.* 10.2. 1–4 http://dx.doi.org/10.1109/IEDM.2013.6724599 (2013).
28. LeCun, Y., Bottou, L., Orr, G. B. & Müller, K.-R. Efficient backprop. *Lect. Notes Comput. Sci.* **7700,** 9–48 (2012).
29. Schiffmann, W., Joost, M. & Werner, R. Optimization of the Backpropagation Algorithm for Training Multilayer Perceptrons https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.6869&rep=rep1&type=pdf (Technical Report, Institute of Physics, University of Koblenz, 1994).
30. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Neural Inf. Process. Systems* **12,** 1097–1105 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge useful discussions with F. Alibart, I. Kataeva, W. Lu, L. Sengupta, S. Stemmer, and E. Zamanidoost. This work was supported by the AFOSR under the MURI grant FA9550-12-1-0038, by DARPA under contract number HR0011-13-C-0051UPSIDE via BAE Systems, and by the DENSO Corporation, Japan.

Author Contributions M.P., F.M.-B., B.D.H., K.K.L., and D.B.S. designed the research. M.P., B.D.H., and G.C.A. performed fabrication and device testing. M.P. and F.M.-B. performed pattern classifier experiments. All authors discussed and interpreted results. M.P., K.K.L., and D.B.S. wrote the manuscript. K.K.L. and D.B.S. advised on all parts of the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.P. (mprezioso@ece.ucsb.edu) and D.B.S. (strukov@ece.ucsb.edu).