nature electronics

ARTICLES

https://doi.org/10.1038/s41928-019-0321-3

Corrected: Author Correction

# Ferroelectric ternary content-addressable memory for one-shot learning

Kai Ni[1]*, Xunzhao Yin[1], Ann Franchesca Laguna[1], Siddharth Joshi[1], Stefan Dünkel[2], Martin Trentzsch[2], Johannes Müller[2], Sven Beyer[2], Michael Niemier[1], Xiaobo Sharon Hu[1] and Suman Datta[1]

**Deep neural networks are efficient at learning from large sets of labelled data, but struggle to adapt to previously unseen data. In pursuit of generalized artificial intelligence, one approach is to augment neural networks with an attentional memory so that they can draw on already learnt knowledge patterns and adapt to new but similar tasks. In current implementations of such memory augmented neural networks (MANNs), the content of a network's memory is typically transferred from the memory to the compute unit (a central processing unit or graphics processing unit) to calculate similarity or distance norms. The processing unit hardware incurs substantial energy and latency penalties associated with transferring the data from the memory and updating the data at random memory addresses. Here, we show that ternary content-addressable memories (TCAMs) can be used as attentional memories, in which the distance between a query vector and each stored entry is computed within the memory itself, thus avoiding data transfer. Our compact and energy-efficient TCAM cell is based on two ferroelectric field-effect transistors. We evaluate the performance of our ferroelectric TCAM array prototype for one- and few-shot learning applications. When compared with a MANN where cosine distance calculations are performed on a graphics processing unit, the ferroelectric TCAM approach provides a 60-fold reduction in energy and 2,700-fold reduction in latency for a single memory search operation.**

Deep learning[1] has made substantial progress in areas such as computer vision[2], speech recognition[3] and machine translation[4]. However, it struggles to adapt to new data—an essential feature of lifelong learning. When a trained deep neural network encounters previously unseen classes, it often fails to generalize from its previous knowledge and must relearn the network parameters to extract relevant information[5]. This means that large numbers of labelled data need to be available for network training. A biological brain, on the other hand, learns at a rapid pace from just a few examples (and sometimes even just one example), and can efficiently generalize from past experiences. As such, learning how to learn (or meta-learning), is a key area of research in machine learning[6,7], which could eventually enable the construction of truly intelligent machines.

One promising approach for implementing meta-learning is a memory augmented neural network (MANN), where features extracted from a neural network can be stored and retrieved from an attentional memory (dynamic random-access memory (DRAM), for example)[8–12]. Recent demonstrations of this approach include differentiable neural computers[8,9], which can learn to construct complex data structures such as graphs and decision trees, answer questions related to data structures and perform one/few-shot learning tasks[10–12]. One/few-shot learning only requires one/a few training examples from a given class to make accurate predictions. The relevant features (such as for a classification task) are extracted from a few training examples and are stored in the network's memory and later retrieved to make predictions (Fig. 1a).

One key function of the memory module is content-based addressing, where, given an input search vector, the distance between the search vector and all the stored vectors is calculated to find the stored vector nearest to the search vector. In the conventional approach, the stored memory vectors (in DRAM) need to be transferred to a compute unit (central processing unit or graphics processing unit (GPU)) to compare distances for a given query, typically via a cosine distance norm (Supplementary Fig. 10a). For example, for a single memory search operation in a memory with $M$ entries, the total number of required operations for calculating the cosine distance on a GPU backed by DRAM include data transfer for $M$ entries, $MD$ multiply operations ($D$ is the dimension of each memory entry) and $M(D-1)$ add operations. As such, energy dissipation and latency limitations with respect to interconnect bandwidth can represent substantial challenges to scaling up MANNs, as it would be expensive to search through a large, conventional memory (Supplementary Fig. 10b).

Ternary content-addressable memories (TCAMs) can identify the stored vector that exactly matches a query. With a TCAM-based approach, given a query feature vector, one can quickly and efficiently search across all the stored vectors (that is network memories)[13]. It performs the search operations directly within the memory itself, thereby eliminating expensive data transfer between the attentional memory and the compute units. That said, conventional TCAMs only perform exact match per search and act as binary comparators (that is, XNOR gates). With binary comparators, to calculate the distance between the search vector and each of the stored vectors, multiple consecutive searches need to be performed[14]. While this approach offers potential advantages compared with a GPU-based distance metric calculation, it could still incur energy and latency overheads as it involves charging/discharging the match lines (MLs) in a TCAM array multiple times (that is, with each search)[14]. It is much more desirable for the TCAM to compute any requisite distance norm with just a single search, while simultaneously computing the degree of match.

[1]University of Notre Dame, Notre Dame, IN, USA. [2]GLOBALFOUNDRIES Fab1 LLC & Co. KG, Dresden, Germany. *e-mail: kni@nd.edu

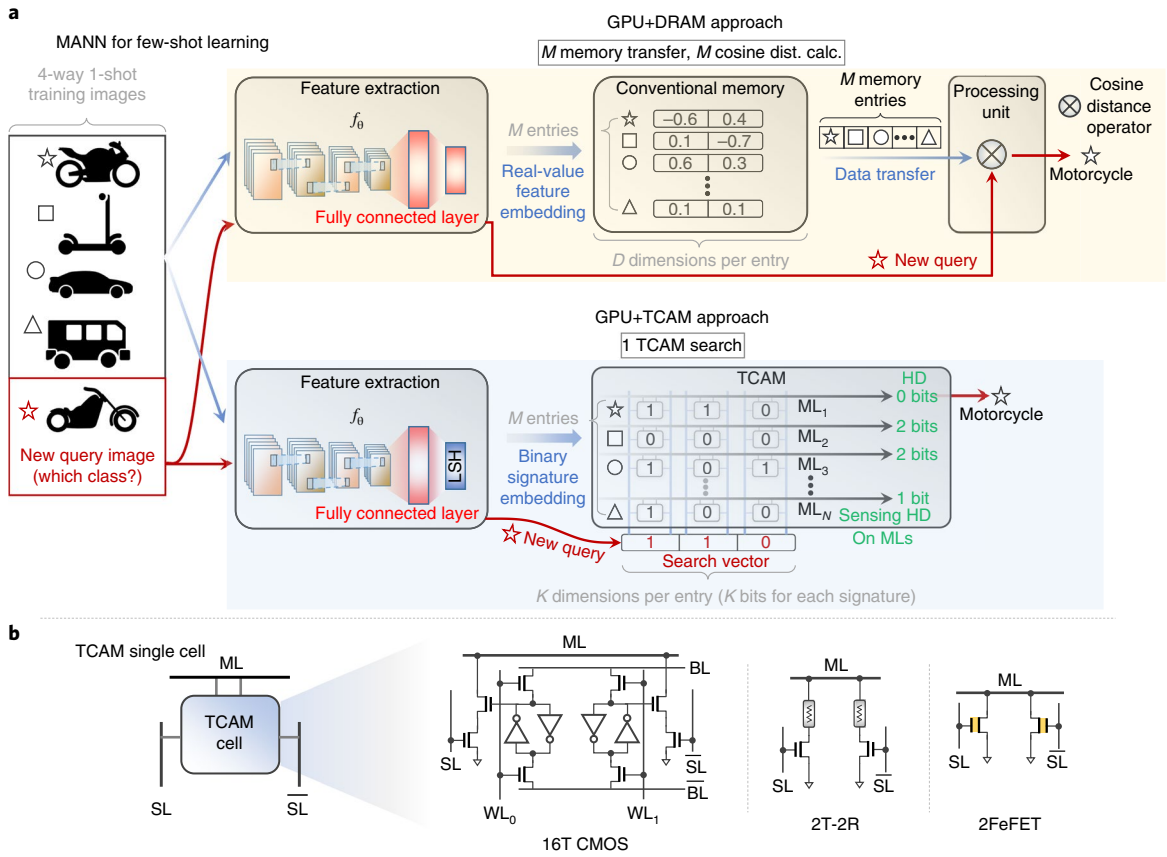



**Fig. 1 | MANN for one-shot learning. a**, The MANN architecture consists of a conventional neural network (for example, the CNN explained in Supplementary Fig. 10a) for feature extraction and an external memory where features are stored and recalled. The external memory is critical for meta-learning, as features extracted for different classes that the trained network has learned before (vehicles in this example) can be stored and retrieved to execute new tasks. When a new example of a previously unseen vehicle is subsequently queried, its extracted feature is compared with all the stored features and the closest feature can be used to predict the class of the new query. MANNs based on conventional memory such as DRAM are difficult to scale up because all the memory entries need to be transferred from the memory to the processing unit to calculate the distance between the query vector and stored vectors. For a memory with *M* entries, the total operation involves *M* entry data transfers and *M* cosine distance calculations. This results in substantial energy and latency penalties due to multiple memory references. TCAMs can act as the external memory to accelerate and scale up MANNs, as TCAMs not only support parallel search operations over all memory entries, but also can compute the Hamming distance (HD) on the match lines (MLs) within a single search operation. To efficiently utilize the Hamming distance compute kernel, we replace the last fully connected layer in the CNN with an LSH layer, which translates the real-valued feature vectors to binary signature vectors. **b**, A single TCAM cell based on CMOS static random-access memory technology requires 16 transistors, occupies a large area and is volatile, and thus incurs a footprint and leakage energy penalty. BL, bit line; SL, search line; WL, write line. TCAMs based on resistive storage elements reduce the cell footprint, but suffer from high write energy, low ON/OFF ratios and variation issues. TCAMs based on two FeFETs have the smallest cell footprint and are non-volatile, and hence achieve the best energy and latency product performance[24]. Vehicle icons in **a** made by Freepik from www.flaticon.com.

As the mismatch between the search vector and stored vector increases (or decreases), an ML discharges faster (or slower)[15,16]. Therefore, it is possible to quantitatively and accurately compute the degree of match by sensing the discharge rates of the MLs. Note that a TCAM only computes the Hamming distance norm between the query and the stored vectors, through direct measurement of the number of mismatched bits. The cosine distance norm calculation, a widely used distance metric in meta-learning applications, is not amenable to TCAM implementations. To address this problem, a locality sensitive hashing (LSH) function[17] can be used to hash real-value feature vectors to a binary signature as an intermediate step (Fig. 1a). A TCAM array can then be used to find the Hamming distance between the binary signatures (Supplementary Fig. 10a). LSH encodes the feature vectors such that similar vectors will be mapped with the same signature. The stored vector with the minimum Hamming distance from the query vector is chosen as the nearest vector. With this approach, TCAMs can be employed to implement the external memory in a MANN, and calculate the distance between the requested search vector and all the memory entries with just a single parallel search across the array (Supplementary Fig. 10b), which will substantially improve both energy and latency in memory search operations.

The use of TCAMs also accelerates the memory update operation (Supplementary Fig. 10c). In a GPU-based MANN backed by DRAM, the content-based memory update operation is inefficient[18,19] (Supplementary Fig. 12). After the memory search operation, which identifies the index of the memory entry that needs to be updated (either the nearest neighbour or the least recently used term; Supplementary Fig. 10c), the GPU needs to locate the memory entry at a given index. This step is computationally expensive, as a GPU (with graphics DRAM) cannot efficiently parallelize write operations to a random address, and therefore suffers from low memory bandwidth use. However, with content-based memory access in TCAMs, the memory entry to be updated is directly

written after a memory search operation, which leads to substantial savings in both energy and latency.

Conventional TCAMs based on complementary metal–oxide–semiconductor (CMOS) static random-access memory technology are volatile, require 16 transistors and occupy a large area (Fig. 1b)[20]. On the other hand, TCAMs based on emerging non-volatile memories (NVMs) are compact and non-volatile. The most common TCAM cell is based on a 2T-2R configuration, which employs two select transistors and two resistive storage elements (based on resistive memory[21,22] or spin-transfer torque magnetic random-access memory[23]). TCAM designs based on resistive storage elements are fully compatible with CMOS processes and consume less area than the 16T TCAMs using conventional CMOS. Still, several challenges persist with TCAMs based on resistive storage elements: leakage power is high due to the low resistance values and the extra peripheral sensing circuitry necessitated by the limited sense margin caused by low resistance ratios ($R_{high}/R_{low}$); write power is high due to the current-driven write mechanism and the need for large access transistors to supply the current; and storage elements must be integrated via the back-end-of-line process, thereby worsening the electrical parasitics.

Alternatively, the recent discovery of ferroelectricity in doped $HfO_2$ has created interest in its integration as a gate dielectric in ferroelectric field-effect transistors (FeFETs), due to its excellent scalability and CMOS compatibility[24]. The FeFET operates by setting the ferroelectric polarization orthogonal to the plane of the semiconductor channel and the gate electrode. The direction of the polarization (pointing towards or away from the channel) sets the FeFET threshold voltage ($V_{TH}$) to be low or high, respectively[25,26]. Unlike current-driven NVMs, the write operation in a FeFET is electric field driven, thereby eliminating the large write currents and resulting in superior energy efficiency. The read operation is aided by the transistor amplification process, thereby allowing extremely fast, reliable and non-destructive access. Recently, we proposed and simulated a compact and energy-efficient TCAM implementation based on two FeFETs (Fig. 1b)[27]. Note that this design is more compact and efficient compared with previously proposed FeFET TCAM designs[26], which are also based on a more idealized single-domain ferroelectric model that is much less representative of the characteristics of FeFETs fabricated in practice.

In this Article, we report an experimental demonstration of a compact and energy-efficient TCAM cell based on FeFETs. We illustrate the capability of the 2FeFET TCAM array in quantitatively sensing the degree of match between the search data and all the stored data in a parallel fashion. Such capability is shown to improve the overall MANN system energy and latency performance, in the context of one-shot and few-shot learning applications.

## 2FeFET TCAM single cell and array

The FeFET used in this work utilizes silicon-doped $HfO_2$ as the ferroelectric gate dielectric and is based on a 28 nm industrial high-$\kappa$ metal gate process[28]. It features a gate stack comprised of poly-Si/2 nm TiN/8 nm doped $HfO_2$/1 nm $SiO_2$ interlayer/p-Si substrate, as shown in Fig. 2a. The device characteristics of an isolated FeFET in response to ±4-V-amplitude, 10-µs-wide write pulses are shown in Fig. 2b. Under the positive gate pulse, the polarization is switched in the direction pointing towards the semiconductor channel, which inverts the channel and sets (erases) the FeFET to a low-$V_{TH}$ state. Similarly, the negative gate pulse flips the polarization direction and sets the device to a high-$V_{TH}$ state. The difference in $V_{TH}$ between the two states represents the device's memory window, which is 1.1 V in this case. The switching dynamics in the FeFET are shown in Fig. 2c, where the memory window is characterized as a function of the write pulse amplitude and pulse width. This highlights the trade-off between the pulse amplitude and the pulse width. For a given memory window, the applied pulse width decreases exponentially as the pulse amplitude increases linearly (Supplementary Fig. 1). This dependence reflects the domain nucleation process involved during the ferroelectric polarization switching[29,30].

The operation of the single FeFET TCAM cell is demonstrated in Fig. 2d. The cell structure makes it natural to utilize the FeFET AND type memory array[31], where the signal lines connecting to the drain and to the source are parallel in a word for the TCAM demonstration. During the cell operation, complementary states are first written into the two FeFETs, and if the search data match with the stored information the ML remains high; otherwise, the ML discharges. In this example, we write the logic '0' state into the TCAM cell by setting the left/right FeFET to a low-$V_{TH}$/high-$V_{TH}$ state, respectively, and write the logic '1' state by setting the left/right FeFET to a high-$V_{TH}$/ low-$V_{TH}$ state. Next, we search logic 0 by applying a low/high pulse amplitude to the left/right FeFET, respectively, and search logic 1 by applying a high/low pulse amplitude to the left/right FeFET, respectively. With the above write and search schemes, when the stored data and search data match (for example the stored datum is logic 0 and the search datum is logic 0), the FeFET with the low-$V_{TH}$ state is cut off as the voltage of the FeFET $V_G = 0$ V. The FeFET with the high-$V_{TH}$ state is also in a cut-off state with $V_G = V_R$ (0.8 V in this case), which is within the memory window. Therefore, the discharge current is minimal and the ML stays high. However, when search data do not match the stored data, the FeFET with the high-$V_{TH}$ state remains OFF but the FeFET with the low-$V_{TH}$ state is turned ON with $V_G = V_R$, which discharges the ML. Cases where the search datum is logic 1 (Supplementary Fig. 2) and a TCAM cell with a ternary don't care state (Supplementary Fig. 3) are also demonstrated here. Thus, the TCAM cell with two FeFETs is fully functional. Unlike two-terminal NVMs—where the write path and the read path are the same—the FeFET benefits from separate ports for read and write operations. In addition, the leakage current of the FeFET with the low-$V_{TH}$ state in the matched cell can be further suppressed by applying a negative search pulse, instead of 0 V as shown here. This further improves the sensing margin[31].

We note that the seemingly slow operation speed (~ms) reported here is an artifact of our current experimental set-up, and not related to the intrinsic speed of the FeFET. As shown in Fig. 2c, FeFETs can be written with ±4 V pulses with 50 ns pulse widths. Write times as low as 10 ns have been reported elsewhere[32]. The search operation is a regular transistor read, which does not involve polarization switching and is intrinsically fast. Therefore, when parasitics are minimized (as in a fully integrated monolithic TCAM sensing circuit) the operation speed of a FeFET TCAM will be fast (~ns), as shown via the circuit simulations based on an experimentally calibrated FeFET compact model[33] (Supplementary Fig. 4). An experimental demonstration of a fully functional 2×2 TCAM array is shown in Fig. 2e. Here we present the full match and the full mismatch cases. During a search operation, the TCAM simultaneously compares the stored data in each row with the search data. Each ML is independent and indicates whether the data stored in that row matches the search data. In the full match case, both MLs remain high as the discharge current is small. In the full mismatch case, both MLs discharge faster than a single cell (Fig. 2d), as the two TCAM cells in a row discharge at the same time, doubling the discharge current.

## Degree of match

Depending on the number of TCAM cells whose stored data match the search data, the ML discharge rate will be different, allowing for the detection of the degree of match using TCAMs. Here, we experimentally demonstrate the direct sensing of the degree of match in a 1×6 TCAM array, as shown in Fig. 3a. (The array size is chosen on the basis of the current structure of our experimental set-up.) A transmission electron microscopy cross-sectional image of a one-dimensional array is also included. Initially, every TCAM

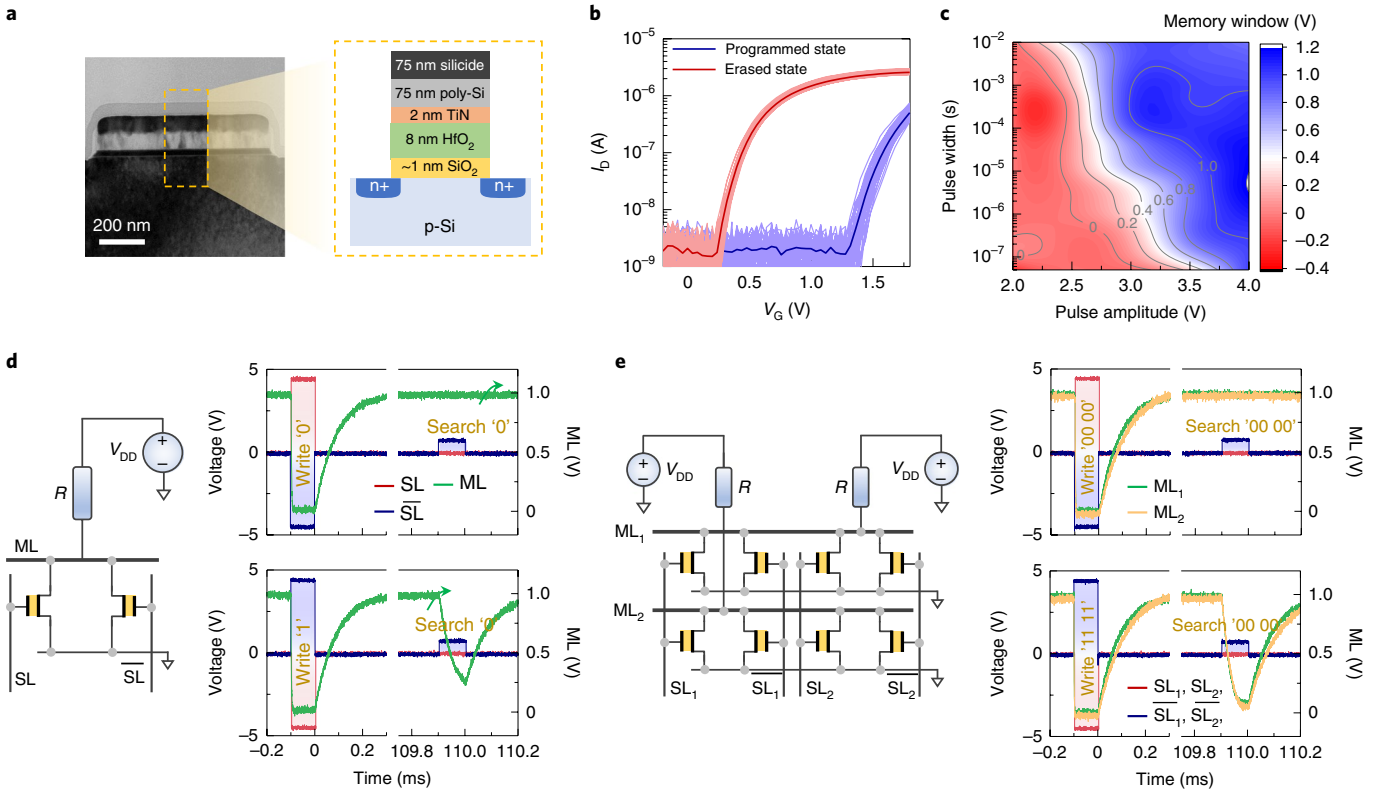



**Fig. 2 | Ferroelectric TCAM cell operation. a**, Transmission electron microscopy cross-section image of a FeFET using a 28 nm CMOS compatible fabrication process. The device gate stack is composed of polycrystalline Si/2 nm TiN/8 nm doped $HfO_2$/1 nm $SiO_2$/p-type Si channel. **b**, FeFET characteristics after −4 V, 10 µs (program) and +4 V, 10 µs (erase) gate pulses. Here 50 full program/erase cycles are presented. The solid line represents the average of the 50 cycles. FeFETs exhibit well controlled cycle-to-cycle variation. Measurement is performed with $V_{DS}$ = 50 mV. **c**, FeFET memory window as a function of applied pulse amplitude and pulse width. **d**, Operation of a single 2FeFET TCAM cell. When the written data match the search data, the ML discharge current is small and the ML voltage stays high; when a data mismatch happens, a large discharge current flows through the low-$V_{TH}$ FeFET and the ML discharges. ($V_{DD}$ = 1.5 V and $R$ = 0.47 MΩ are used in this paper.) **e**, 2 × 2 TCAM array operation. The two MLs are independent of each other. In the full match case the two MLs remain high, while in the full mismatch case the two MLs discharge at a rate higher than that for the single TCAM cell.

cell is written with logic 0. Then, a set of search patterns, that is, 000000, 000001, 000011, 000111, 001111, 011111, 111111, is applied sequentially to the array. Such write and search operations allow the array to sense the total number of mismatched bits between the search data and stored data in a sequential manner. Figure 3b shows the measured device characteristics for six FeFETs in the array. A well controlled distribution of both program and erase states and a wide opening of the memory window is observed. Cycle-to-cycle and device-to-device variations are summarized in Supplementary Fig. 5, highlighting well controlled distribution statistics for the fabricated FeFETs.

The measured ML transient waveforms with an increasing number of mismatched bits (from zero to six) are shown in Fig. 3c. As expected, the discharge rate of the ML increases with the number of mismatched bits. We model the ML waveform with an exponential decay function, $1-\exp(-t/\tau)$. The discharge rate, $1/\tau$, shown in Fig. 3d, exhibits a linear dependence on the number of mismatched bits. This is because the discharge current increases linearly with the number of mismatched bits. Similarly, the extracted ML voltage depends linearly on the number of mismatched bits. These results demonstrate, for the first time, that the FeFET TCAMs can successfully detect the number of mismatched bits, which is essentially the Hamming distance between the search vector data and the stored vector data. Additionally, we designed and simulated a sense amplifier to detect the degree of match (Supplementary Fig. 6 and Supplementary Fig. 7) for fully integrated TCAM implementation.

The effects of TCAM array size and FeFET device-to-device variation on sensing the degree of match are shown in Supplementary Figs. 8 and 9. As the TCAM array size scales up, one important question is how large a Hamming distance between the query vector and stored vectors must be sensed. For one-shot learning, given a query, the nearest neighbour in the TCAM array needs to be identified. Hence, it is not necessary to search for entries whose Hamming distance is sufficiently large that they are too far away and irrelevant to the query. Therefore, we studied Hamming distances up to eight bits, as a representative case study, for different array sizes. Our initial analysis suggests that the array size does not have an overly substantial effect on the detection, as shown in Supplementary Fig. 8a. Additionally, when taking FeFET variation into consideration, all possible Hamming distances (zero to eight bits) can be reliably detected in an 1 × 8 TCAM array. This functionality still holds after considering the metal line parasitic resistances, as shown in Supplementary Fig. 9. Extending beyond 1 × 8 arrays is likely with future FeFET process optimization. Moreover, one advantage of Hamming distance calculations is that they can be performed in segments. This allows a large array to be divided into several small subarrays so that all the degrees of match can be detected. Therefore, even a current FeFET TCAM array could support MANN applications.

## Ferroelectric TCAM augmented neural network

To further demonstrate the utility of the ferroelectric TCAM hardware, we evaluate the performance of a MANN with the 2FeFET

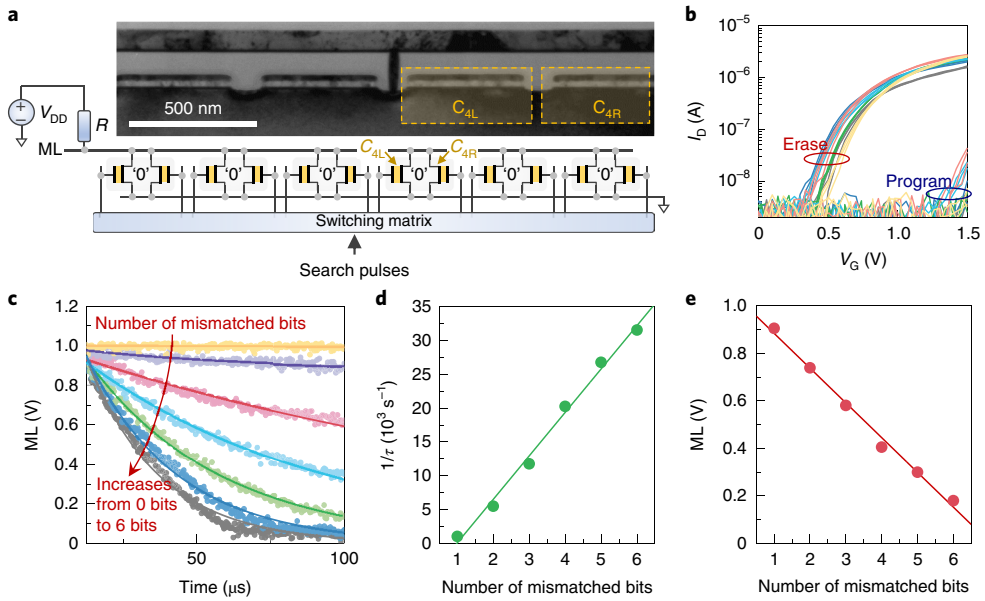



**Fig. 3 | Degree of match measurement using ferroelectric TCAM array. a**, Transmission electron microscopy cross-section of FeFETs in an array and schematic of experimental set-up for the degree of match measurement. Each TCAM cell is initially written with bit 0, then search data with successively increasing numbers of 1 bits are applied to monitor the discharge characteristics of the ML. **b**, FeFET characteristics after −4 V, 10 μs (program) and +4 V, 10 μs (erase) gate pulses. Each colour represents one device and a total of six devices are shown. For each FeFET, three program/erase cycles are presented. The FeFETs exhibit well controlled device-to-device variation. **c**, The ML discharge rate increases with the number of mismatched bits due to the increase of the discharge current. The circles represent the measured transient waveform and the lines represent the compact model with an exponential decay function, $1-\exp(-t/\tau)$. **d**, $1/\tau$ exhibits a linear dependence on the number of mismatched bits. **e**, The ML voltage extracted at 50 μs exhibits a linear dependence on the number of mismatched bits.

TCAM array serving as the external memory. One-shot and few-shot learning tasks are used as a case study, where a network learns from one or a few labelled examples per class[10,11], a key challenge in machine learning for artificial general intelligence. Again, MANNs represent a promising solution to this challenge by storing learned features from new classes in an external memory that can later be retrieved for prediction[10,11]. We evaluate the performance of the ferroelectric TCAM augmented neural network for one- and few-shot learning applications.

In our study, we consider one- and few-shot learning in the context of the Omniglot dataset[34]—a commonly used benchmark for few-shot learning[12]—that contains 1,623 characters from 50 different alphabets. A total of 1,200 characters are used for training while the remaining characters are used for testing. The Omniglot dataset has a large number of classes (1,623 classes) and only 20 examples for each class. In an *N*-way, *K*-shot learning application, the neural network is shown a set of supporting images from *N* classes, with *K* examples of each class. It is then shown a new query image (from amongst the *N* classes) and must identify which supporting image the query image is most similar to (that is, which class it belongs to). For example, in a five-way, one-shot learning task, a network will be shown five new classes, with just one example from each. It will then be shown a new example pertaining to one of the five classes and be expected to identify it.

In the case of a MANN implemented with a GPU backed by DRAM, a regular network such as a convolutional neural network (CNN) first extracts relevant features from an image (Supplementary Fig. 10a), which are then stored in memory and later transferred from the memory to the compute unit for cosine distance calculations. In the GPU backed by TCAM approach, we replace the last fully connected layer in a CNN with an LSH function layer (Supplementary Fig. 10a). Note that a typical neural network usually ends with a softmax layer, but here we replace this with an LSH function and the TCAM module. The LSH layer is used to convert features from a real-valued feature vector to a binary signature (as illustrated in the simplified example in Fig. 4a). As the LSH and the original fully connected layer have similar computational demands, the replacement does not incur additional overhead in either storage or computation. LSH with random projections is used such that feature vectors (for example, images) associated with the same class have similar binary signatures (for example, the star-shaped objects in Fig. 4a should share similar binary signatures). Thus, a greater (or smaller) Hamming distance between the binary signatures implies less (or more) similarity between the features represented by the signatures. By storing the binary signatures in the TCAM, one can search in parallel for the supporting class that is closest to a given query. Usually the number of LSH hashing planes (*K* in Fig. 4a) is a hyperparameter and is tuned until its further increase does not further improve system accuracy. Figure 4b illustrates the classification accuracy of one-shot/few-shot learning tasks (for the Omniglot dataset) assuming both a cosine similarity distance metric (with comparisons performed on a GPU) and a MANN that leverages the aforementioned distance compute kernel implemented in the TCAM hardware. The TCAM-based MANN exhibits classification accuracies that approach (and sometimes match) those obtained with the conventional cosine distance calculation when implemented on a GPU backed by external DRAM. Note that with the LSH approach it is also possible to apply the binary CAM cells for Hamming distance calculation. The ternary state of TCAM, the don't care state, is not necessary to realize the Hamming distance calculation function. However, the ferroelectric TCAM cell design incurs no additional overhead compared with the binary CAM cell, as they are exactly the same. Additionally, there are LSH algorithms, such as the ternary LSH, which require the third state and perform better in separating different points and mapping similar points together than simple binary LSH[35].

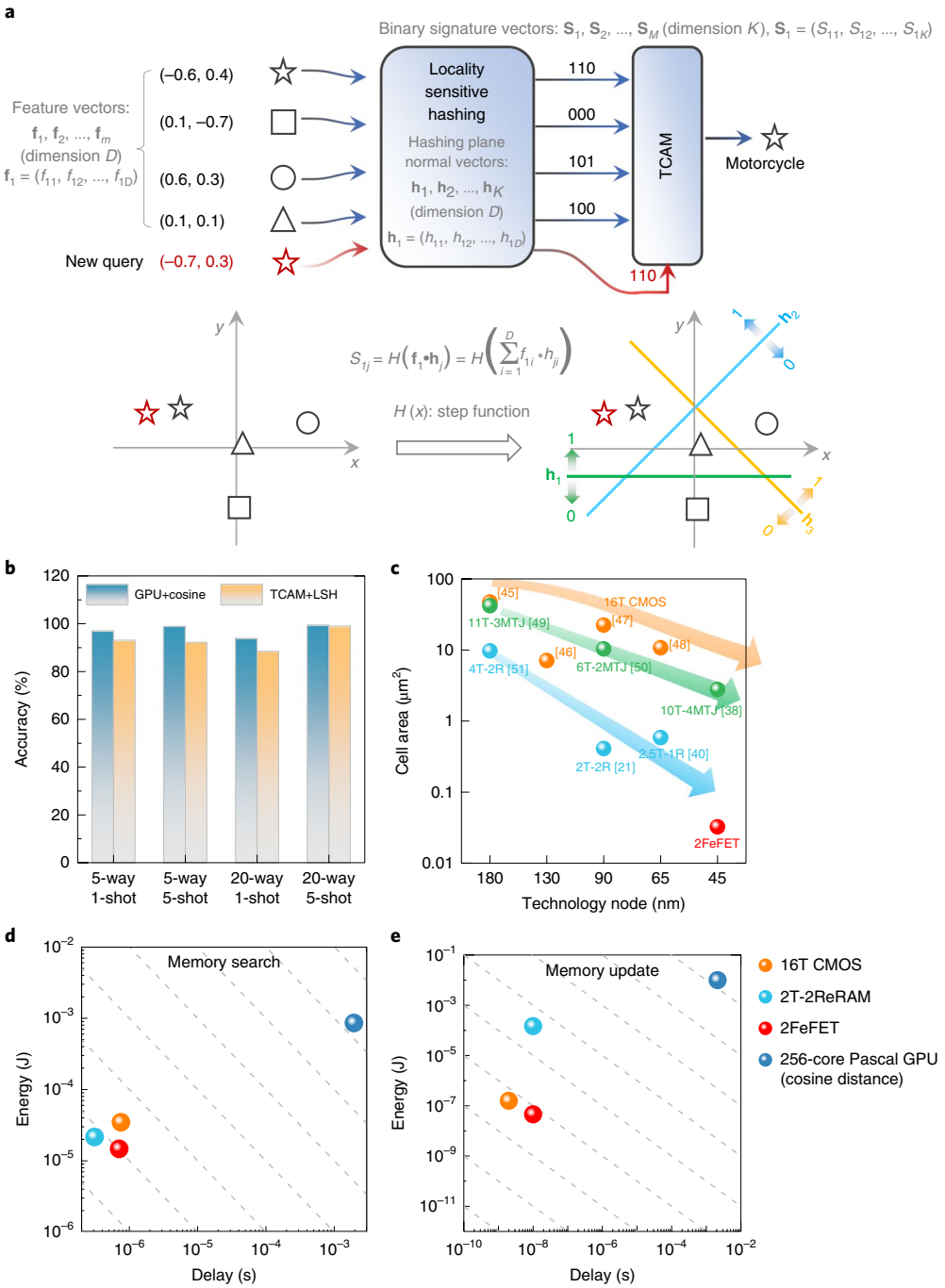
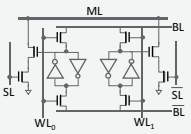
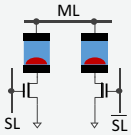
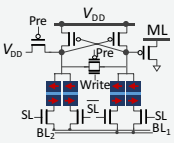
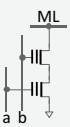
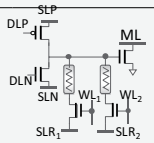
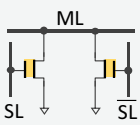



**Fig. 4 | One- and few-shot learning with TCAM. a**, LSH function to map a feature vector from a real-valued space to a binary signature space. *M* feature vectors, where each feature vector has a dimension of *D*, are fed to the LSH layer. For illustrative purposes, this example assumes a two-dimensional feature space ($D=2$). Four support feature vectors ($M=4$) are distributed in this space. An LSH function uses the hashing planes with normal vectors $\mathbf{h}_1$, $\mathbf{h}_2$, …, $\mathbf{h}_K$. In this example, three random planes are shown, $K=3$, to partition the feature space. Therefore, each feature vector is encoded into *K* binary signature bits depending on the location of the feature vector with respect to the *K* hashing planes. Each feature vector is encoded with three signature bits, where each bit is assigned value 1 if the feature vector is above the hyperplane and 0 otherwise. For example, the star is encoded as 110 as it sits above the planes $\mathbf{h}_1$ and $\mathbf{h}_2$ and below the plane $\mathbf{h}_3$. **b**, The classification accuracy of one/few-shot learning on the Omniglot dataset. **c**, A TCAM cell area with different technology nodes for CMOS and different NVM technologies. The 2FeFET TCAM is the most compact. Data from refs. [21,38,40,45–51]. **d**,**e**, Benchmark of energy–latency product metric for single memory search and update operations, respectively. For the GPU case, the energy and delay metric involves the memory access operation and the cosine distance calculation. 2FeFET TCAM exhibits 60-fold and 2,700-fold reductions in energy and delay for a single search operation and $2\times10^5$-fold and $2.2\times10^5$-fold reductions in energy and delay for a single memory update operation, compared with the conventional GPU and DRAM implementation.

The above results suggest that using the hashing function in the GPU allows us to harness the efficient distance metric calculations in TCAM arrays, which can overcome fundamental data transfer bottlenecks associated with traditional memories, without substantial degradations in accuracy. More quantitatively, Fig. 4c benchmarks the area of a single TCAM cell implemented with CMOS

**Table 1 | Performance comparison of TCAM implementations built with various technologies**

| | 16T CMOS | 2T-2PCM (ref. [21]) | 10T-4MTJ (ref. [38]) | 2Flash (ref. [39]) | 2.5T-1ReRAM (ref. [40]) | 2FeFET (this work) |
|---|---|---|---|---|---|---|
| |  |  |  |  |  |  |
| Technology node (nm) | 45 | 45 | 45 | 45 | 45 | 45 |
| Cell area (μm²)[a] | 1.12 (7.5×) | 0.41 (2.7×) | 2.78 (18.5×) | 0.30 (2×) | 0.28 (1.8×) | 0.15 (1×) |
| Non-volatility | No | Yes | Yes | Yes | Yes | Yes |
| $R_{ON}/R_{OFF}$ | ~$10^6$ | ~100 | ~2.5 | ~$10^6$ | ~100 | ~$10^4$ |
| Search voltage (V) | 1 | 1 | 1 | 1.1 | 1 | 1 |
| Search delay[b] (ps) | 582 (1.7×) | 155 (0.44×) | 1,000 (2.9×) | 679 (2×) | 155 (0.44×) | 355 (1×) |
| Search energy (fJ per bit per search) | 1.0 (2.4×) | 0.64 (1.6×) | 40.5 (101.2×) | 0.6 (1.5×) | 0.71 (1.8×) | 0.4 (1×) |
| Normalized EDP | 4.1× | 0.7× | 293.6× | 3× | 0.8× | 1× |
| Write scheme | Voltage driven | Current driven | Current driven | Current driven | Current driven | Voltage driven |
| Write voltage (V) | 1 | 2.5 | 0.6–0.8 | 10–20 | Set: 1.8<br>Reset: 1.2 | ±4 |
| Write delay[c] | <2 ns | ~10 ns (ref. [41]) | ~10 ns (ref. [42]) | 0.1–1 ms | ~10 ns (ref. [43]) | ~10 ns |
| Write energy (fJ per bit) | 4.8 (3.5×) | ~4,500 (3,225×) | 870 (626×) | >98,000 (ref. [44])<br>(>70,000×) | ~720 (514×) | 1.4 (1×) |

[a]The cell area is estimated from the layout. [b]Search delay obtained through SPICE simulation of a 64 × 64 array. [c]Optimistic write delay reported in the literature. The values in parentheses show the ratio of each TCAM technology normalized to ferroelectric TCAM.

as well as other NVM technologies. As a ferroelectric TCAM cell constitutes two FeFETs, the result is an extremely dense structure, thereby allowing for large and dense TCAM arrays. When a 64×64 FeFET TCAM array is employed to compute the distances between one search vector and all the stored vectors in a MANN for one-shot learning, we observe a speedup of 2,700-fold and energy savings of 60-fold in a single memory search operation (Fig. 4d) when compared with a cosine distance-based calculation assuming a GPU backed by conventional DRAM. In addition, a speedup of $2\times10^5$-fold and energy savings of $2.2\times10^5$-fold are observed for a single memory entry update (Fig. 4e), as it is inefficient for a GPU to write to a single, random address. These savings originate, in part, from the improvement at the architecture level via (1) the replacement of conventional DRAM with a TCAM array and (2) the utility of the LSH-based mapping, as suggested by the benchmarking results shown in Supplementary Fig. 13, which reports 24-fold and 2,582-fold reductions in energy and latency, respectively, for a memory search operation when a 16T CMOS TCAM replaces the DRAM. At the cell level, the 2FeFET TCAM achieves the best energy–delay product for both the memory search and the memory update operation among the various TCAM implementations. As shown in Supplementary Fig. 13, replacing 16T CMOS TCAMs with 2FeFET TCAMs further reduces the latency and energy by 1.1-fold and 2.4-fold, respectively, during the memory search operation.

MANNs are ultimately intended for implementing lifelong learning. Hence, some memory entries may not be updated as frequently as others, but must remain in the memory. Unlike FeFET-based TCAMs, 16T CMOS TCAMs are volatile, and hence consume large amounts of leakage power. Resistive memory typically suffers from low $R_{OFF}$ to $R_{ON}$ ratios, and low OFF state resistance compared with FeFETs, which limit its ability to accurately sense the degree of match. Therefore, resistive-memory-based TCAMs usually require additional encoding schemes and peripheral circuitry to measure the degree of match[21]. As summarized in Table 1, which compares ferroelectric TCAM performance with all the other technologies, the compact, energy-efficient and non-volatile 2FeFET TCAM arrays represent an ideal technology platform to implement the attentional memory module in MANNs.

Although the endurance of current $HfO_2$ FeFETs may be degraded after $10^5$ cycles (ref. [24]) (due to extrinsic charge trapping in the $HfO_2$), it still might suffice for the inference applications for one-shot learning, as memory write operations may not be as frequent. In fact, writes are only necessary when a new class must be added to the TCAM. Moreover, recent work suggests that the intrinsic endurance property of ferroelectric $HfO_2$ material could be extended to $10^{12}$ cycles (ref. [25]), suggesting that the $HfO_2$ FeFET endurance can potentially be improved by suppressing the charge trapping through process optimization. Meanwhile, novel ferroelectric memory device design can also be pursued to potentially overcome this endurance bottleneck[36].

Finally, we note that there are many flavours of model-based meta-learning approaches where a query is compared with models stored in memory[14]; the FeFET TCAM-based distance compute kernels described here would be equally amenable to these models as well.

To evaluate the overall system-level performance improvements when ferroelectric TCAMs are employed as external memory in a MANN, we also consider the impact of the neural network itself (that is, $f(\theta)$ in Fig. 1a). Supplementary Fig. 11 shows the decomposition of the execution time for the inference process in a GPU with a conventional DRAM memory. It shows the decomposition of various operations in the memory module, including the memory search and memory update operation. Using Amdahl's law, the overall system improvement can be calculated (Supplementary Fig. 13). For inference, the overall energy and latency improvement

of the entire MANN (the neural network and the memory module) is 4.4-fold and 4.5-fold respectively compared with the Pascal GPU implementation on Jetson TX2 (ref. [37]) hardware, as shown in Supplementary Fig. 14. We note that further improvements would be possible as (1) the MANN is scaled up to tackle more complex meta-learning tasks that demand more memory operations (that is, the percentage of computation performed by the memory network increases) and (2) more efficient neural network accelerators are realized (that is, to perform computations for $f(\theta)$, as our approach is orthogonal to this).

## Conclusions

We implement the external memory in a MANN for one- and few-shot learning with ferroelectric TCAMs, by leveraging its inherent massive parallelism and in-memory computing capability. We experimentally demonstrate a compact and energy-efficient TCAM design that consists of two FeFETs using an industrial 28 nm high-$\kappa$ metal gate FeFET process. The FeFET device technology enables the most high-density, energy-efficient and non-volatile TCAMs demonstrated to date that are suitable for MANN implementation. The ability to calculate Hamming distances between a search vector and stored vectors in a massively parallel fashion by sensing the discharge rate of the MLs is demonstrated. We show that, by translating the real-value feature space to a binary signature space through an LSH function, the Hamming distance calculation functionality is sufficient for acceleration of search operations required in a MANN. As such, the 2FeFET TCAM provides a promising emerging hardware platform towards implementing large-scale MANNs that serve as building blocks for future intelligent lifelong learning machines.

## Methods

**Device fabrication.** In this paper, the fabricated FeFET features a polycrystalline Si/TiN (2 nm)/doped $HfO_2$ (8 nm)/$SiO_2$ (1 nm)/p-Si gate stack. The devices were fabricated using a 28 nm node gate-first high-$\kappa$ metal gate CMOS process on 300 mm silicon wafers. The ferroelectric gate stack process module starts with growth of a thin $SiO_2$ based interfacial layer, followed by the deposition of 8-nm-thick doped $HfO_2$. A TiN metal gate electrode is deposited using physical vapour deposition, and on top of this the poly-Si gate electrode is deposited. The source and drain $n^+$ regions were obtained by phosphorus ion implantation, and were then activated by a rapid thermal annealing at approximately 1,000 °C. This step also results in the formation of the ferroelectric orthorhombic phase within the doped $HfO_2$. All the devices electrically characterized have the same gate length and width dimensions of 500 nm × 500 nm.

**Electrical characterization.** The FeFET device characterization was performed with a Keithley 4200-SCS semiconductor parameter analyser. Two 4225-PMUs (pulse measurement units) were utilized to make the pulsed current–voltage measurement. In the experiment, program and erase pulses were applied and the pulsed $I_D$–$V_G$ ($I_D$, drain current; $V_G$, gate voltage) measurement was performed. The total sweep duration is 5 ms. Note that, to minimize the charge trapping effects on the sensing of the programmed or erased state of the device, we inserted a delay of 100 ms between the measurement and the write pulses to allow a full trapped charge release. For the pulsed measurements, the current resolution is close to 1 nA in our set-up.

The TCAM characterization was performed using a monolithic FeFET chip and an externally connected resistor. We connected the TCAM cell with an external resistor on a breadboard. Input pulses for memory writes and searches were generated with an Agilent 81150A arbitrary function generator. A 1.5 V amplitude $V_{DD}$ supply was provided through a Keithley 4200 SMU (source measurement unit). The TCAM ML voltage transient was sampled through an Agilent DSO9104A digital oscilloscope. In this experiment, a 0.47 MΩ resistor was used. All the write pulses and search pulses have a pulse width of 100 µs due to the large parasitics in our set-up. In a fully integrated TCAM, the operation speed will greatly improve, as shown in the single-FeFET measurement (successful write under 50 ns, ±4 V) in Fig. 2c. In the experiment, +4.5 V erase pulses and −4.5 V program pulses were applied. For the TCAM cell and array demonstration shown in Fig. 2, the search pulse amplitude was set at 0.8 V. However, for the degree of matching measurement (shown in Fig. 3) a search amplitude of 0.7 V was used to avoid early saturation of the ML voltage.

## Data availability

The data that support the plots within this paper and other findings of this study are available from the corresponding author on reasonable request.

Received: 19 December 2018; Accepted: 30 September 2019;
Published online: 18 November 2019

## References

1. LeCun, Y., Bengio, Y. & Hinton, G. E. Deep learning. *Nature* **521**, 436–444 (2015).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems 25 (NIPS 2012)* (eds Pereira, F. et al.) 1090–1098 (Neural Information Processing Systems Foundation, 2012).
3. Graves, A., Mohamed, A. R. & Hinton, G. E. Speech recognition with deep recurrent neural networks. In *Proc. 2013 IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)* 6645–6649 (IEEE, 2013).
4. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Proc. Advances in Neural Information Processing Systems 27 (NIPS 2014)* (eds Ghahramani, Z. et al.) 3104–3112 (Neural Information Processing Systems Foundation, 2014).
5. McCloskey, M. & Cohen, N. J. Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv.* **24**, 109–165 (1989).
6. Youssef, D. & Vilalta, R. A perspective view and survey of meta-learning. *Artif. Intell. Rev.* **18**, 77–95 (2002).
7. Lemke, C., Budka, M. & Gabrys, B. Meta-learning: a survey of trends and technologies. *Artif. Intell. Rev.* **44**, 117–130 (2015).
8. Graves, A., Wayne, G. & Danihelka, I. Neural Turing machines. Preprint at http://arxiv.org/abs/1410.5401 (2014).
9. Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471–476 (2016).
10. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. & Lillicrap, T. Meta-learning with memory-augmented neural networks. *Proc. Machine Learning Res.* **48**, 1842–1850 (2016).
11. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K. & Wierstra, D. Matching networks for one shot learning. In *Proc. Advances in Neural Information Processing Systems 29 (NIPS 2016)* (eds Lee, D. D. et al.) 3637–3645 (Neural Information Processing Systems Foundation, 2016).
12. Kaiser, L., Nachum, O., Roy, A. & Bengio, S. Learning to remember rare events. In *Proc. Int. Conference on Learning Representations* (2017).
13. Karam, R. et al. Emerging trends in design and applications of memory-based computing and content-addressable memories. *Proc. IEEE* **103**, 1311–1330 (2015).
14. Laguna, A. F., Niemier, M. & Hu, X. S. Design of hardware friendly memory enhanced neural networks. In *2019 Design, Automation and Test in Europe Conference & Exhibition (DATE)* 1583–1586 (IEEE, 2019).
15. Imani, M., Patil, S. & Rosing, T. S. Approximate computing using multiple-access single-charge associative memory. *IEEE Trans. Emerg. Top. Comput.* **6**, 305–316 (2018).
16. Pagiamtzis, K., Azizi, N. & Najm, F. N. A soft-error tolerant content-addressable memory (CAM) using an error-correcting-match scheme. In *IEEE Custom Integrated Circuits Conference 2006* 301–304 (IEEE, 2007).
17. Andoni, A. & Indyk, P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)* 459–468 (IEEE, 2006).
18. Lai, Z., Luo, Q. & Jia, X. Revisiting multi-pass scatter and gather on GPUs. In *Proceedings of the 47th Int. Conference on Parallel Processing* 25 (ACM, 2018).
19. He, A., Govindaraju, N. K., Luo, Q. & Smith, B. Efficient gather and scatter operations on graphics processors. In *SC '07 Proceedings of the ACM/IEEE Conference on Supercomputing* 46 (ACM, 2007).
20. Nii, K. et al. A 28nm 400 MHz 4-parallel 1.6Gsearch/s 80Mb ternary CAM. In *2014 IEEE Int. Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* 240–242 (IEEE, 2014).
21. Li, J. et al. 1 Mb 0.41 µm² 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing. *IEEE J. Solid-State Circuits* **49**, 896–907 (2014).
22. Imani, M., Rahimi, A., Kong, D., Rosing, T. & Rabaey, J. M. Exploring hyperdimensional associative memory. In *2017 IEEE Int. Symposium on High Performance Computer Architecture (HPCA)* 445–456 (IEEE, 2017).
23. Matsunaga, S. et al. Fine-grained power-gating scheme of a metal-oxide-semiconductor and magnetic-tunnel-junction-hybrid bit-serial ternary content addressabl ememory. *Jpn. J. Appl. Phys.* **49**, 04DM05 (2010).
24. Müller, J. et al. Ferroelectricity in $HfO_2$ enables nonvolatile data storage in 28 nm HKMG. In *2012 Symposium on VLSI Technology (VLSIT)* 25–26 (IEEE, 2012).
25. Ni, K. et al. Critical role of interlayer in $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric FET nonvolatile memory performance. *IEEE Trans. Electron Devices* **65**, 2461–2469 (2018).
26. Yin, X., Niemier, M. & Hu, X. S. Design and benchmarking of ferroelectric FET based TCAM. In *DATE '17 Proceedings of the Conference on Design, Automation and Test in Europe* 1448–1453 (European Design and Automation Association, 2017).

27. Yin, X., Ni, K., Reis, D., Datta, S., Niemier, M. & Hu, X. S. An ultra-dense 2FeFET TCAM design based on a multi-domain FeFET model. *IEEE Trans. Circuits Syst. II.* **66**, 1577–1581 (2018).
28. Trentzsch, M. et al. A 28 nm HKMG super low power embedded NVM technology based on ferroelectric FETs. In *2016 IEEE Int. Electron Devices Meeting (IEDM)* 294–297 (IEEE, 2017).
29. Mulaosmanovic, H., Mikolajick, T. & Slesazeck, S. Accumulative polarization reversal in nanoscale ferroelectric transistors. *ACS Appl. Mater. Interfaces* **10**, 23997–24002 (2018).
30. Mulaosmanovic, H. et al. Switching kinetics in nanoscale hafnium oxide based ferroelectric field-effect transistors. *ACS Appl. Mater. Interfaces* **9**, 3792–3798 (2017).
31. Ni, K. et al. Write disturb in ferroelectric FETs and its implication for 1T-FeFET AND memory arrays. *IEEE Electron Device Lett.* **39**, 1656–1659 (2018).
32. Dünkel, S. et al. A FeFET based super-low-power ultra-fast embedded NVM technology for 22 nm FDSOI and beyond. In *2017 IEEE Int. Electron Devices Meeting (IEDM)* 485–488 (IEEE, 2018).
33. Ni, K., Jerry, M., Smaith, J. A. & Datta, S. A circuit compatible accurate compact model for ferroelectric-FETs. In *2018 IEEE Symposium on VLSI Technology* 131–132 (IEEE, 2018).
34. Lake, B. M. et al. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
35. Shinde, R., Goel, A., Gupta, P. & Dutta, D. Similarity search and locality sensitive hashing using ternary content addressable memories. In *SIGMOD '10 Proc. 2010 ACM SIGMOD Int. Conference on Management of Data* 375–386 (ACM, 2010).
36. Ni, K. et al. SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications. In *2018 IEEE Int. Electron Devices Meeting (IEDM)* 296–299 (IEEE, 2019).
37. Franklin, D. NVIDIA Jetson TX2 delivers twice the intelligence to the edge. *NVIDIA Developer Blog* https://devblogs.nvidia.com/jetson-tx2-delivers-twice-intelligence-edge/ (2017).
38. Song, B. et al. A 10T-4MTJ nonvolatile ternary CAM cell for reliable search operation and a compact area. *IEEE Trans. Circuits Syst. II* **64**, 700–704 (2017).
39. Fedorov, V. V., Abusultan, M. & Khatri, S. P. An area-efficient ternary CAM design using floating gate transistors. In *2014 IEEE 32nd Int. Conference on Computer Design (ICCD)* 55–60 (IEEE, 2014).
40. Lin, C. C. et al. A 256b-wordlength ReRAM-based TCAM with 1ns search-time and 14× improvement in word length-energy efficiency-density product using 2.5T1R cell. In *2016 IEEE Int. Solid-State Circuits Conference (ISSCC)* 136–138 (IEEE, 2016).
41. Ahn, S. J. et al. Highly manufacturable high density phase change memory of 64Mb and beyond. In *IEDM Technical Digest. IEEE Int. Electron Devices Meeting, 2004* 907–910 (IEEE, 2005).
42. Lin, C. J. et al. 45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell. In *2009 IEEE Int. Electron Devices Meeting (IEDM)* 279–282 (IEEE, 2010).
43. Govoreanu, B. et al. 10×10nm$^2$ Hf/HfO$_x$ crossbar resistive RAM with excellent performance reliability and low-energy operation. In *2011 Int. Electron Devices Meeting* 729–732 (IEEE, 2012).
44. Dong, Q. et al. A 1Mb embedded NOR flash memory with 39μW program power for mm-scale high-temperature sensor nodes. In *2017 IEEE Int. Solid-State Circuits Conference (ISSCC)* 198–200 (IEEE, 2017).
45. Matsunaga, S. et al. A 3.14 μm$^2$ 4T-2MTJ-cell fully parallel TCAM based on nonvolatile logic-in-memory architecture. In *2012 Symposium on VLSI Circuits (VLSIC)* 44–45 (IEEE, 2012).
46. Roth, A., Foss, D., McKenzie, R. & Perry, D. Advanced ternary CAM circuits on 0.13 μm logic process technology. In *Proc. IEEE 2004 Custom Integrated Circuits Conference* 465–468 (IEEE, 2004).
47. Choi, S., Sohn, K. & Yoo, H. J. A 0.7-fJ/bit/search 2.2-ns search time hybrid-type TCAM architecture. *IEEE J. Solid-State Circuits* **40**, 254–260 (2005).
48. Huang, P. T. & Hwang, W. A 65 nm 0.165 fJ/bit/search 256 × 144 TCAM macro design for IPv6 lookup tables. *IEEE J. Solid-State Circuits* **46**, 507–519 (2011).
49. Xu, W., Zhang, T. & Chen, Y. Design of spin-torque transfer magnetoresistive RAM and CAM/TCAM with high sensing and search speed. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **18**, 66–74 (2010).
50. Matsunaga, S. et al. Fully parallel 6T-2MTJ nonvolatile TCAM with single-transistor-based self match-line discharge control. In *2011 Symposium on VLSI Circuits—Digest of Technical Papers* 298–299 (IEEE, 2011).
51. Huang, L. Y. et al. ReRAM-based 4T2R nonvolatile TCAM with 7x NVM-stress reduction, and 4x improvement in speed-wordlength-capacity for normally-off instant-on filter-based search engines used in big-data processing. In *2014 Symposium on VLSI Circuits Digest of Technical Papers* 298–299 (IEEE, 2014).


## Acknowledgements

This work was supported in part by ASCENT, one of six centres in JUMP, sponsored by DARPA and the Semiconductor Research Corporation (SRC).

## Author contributions

M.N., X.S.H. and S.D. proposed and supervised the project. S.D., M.T., J.M. and S.B. fabricated the devices. K.N. performed the experiment. X.Y. and K.N. conducted the circuit simulations and variation analysis. A.F.L. and S.J. proposed the LSH encoding. A.F.L. performed architecture-level benchmarking. K.N., X.Y., A.F.L., M.N., X.S.H. and S.D. analysed the data. K.N. wrote the paper. All authors contributed to discussions on the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41928-019-0321-3.

**Correspondence and requests for materials** should be addressed to K.N.

# Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com