One-Shot Learning with Memory-Augmented Neural Networks Using a 64-kbit, 118 GOPS/W RRAM-Based Non-Volatile Associative Memory

Haitong Li[*], Wei-Chen Chen, Akash Levy, Ching-Hua Wang, Hongjie Wang, Po-Han Chen, Weier Wan, H.-S. Philip Wong[#], Priyanka Raina[&]

Department of Electrical Engineering and Stanford SystemX Alliance, Stanford University, CA, USA

Email: [*]haitongl@stanford.edu; [#]hspwong@stanford.edu; [&]praina@stanford.edu

Abstract

Learning from a few examples (one/few-shot learning) on the fly is a key challenge for on-device machine intelligence. We present the first chip-level demonstration of one-shot learning using a 2T-2R resistive RAM (RRAM) non-volatile associative memory (AM) as the backend of memory-augmented neural networks (MANNs). The 64-kbit fully integrated RRAM-CMOS AM core (0.2 mm^2 at 40 nm node) enables long-term feature embedding and retrieval, demonstrated in a challenging 32-way one-shot learning task using Omniglot dataset. Using only one example per class for 32 unseen classes during on-chip learning, our AM chip achieves ~72% measured inference accuracy on Omniglot as the first chip accuracy report compared to software accuracy (~82%), while reaching 118 GOPS/W for in-memory L1 distance computation and prediction.

Introduction

On-device machine intelligence requires continuous real-time learning of never-before-seen data/events. Memory-augmented neural networks (MANNs) [1] aim to address this demand by utilizing an explicit associative memory to augment the feature learning capabilities of neural networks (NNs) with scarce data. A MANN consists of a frontend neural feature extractor (such as a convolutional neural network) and a backend associative memory (AM). The backend AM is where learning occurs in the form of new feature embedding (Fig. 1), and where inference occurs as similarity-based retrieval. Here, we develop a native hardware realization for MANN's backend using an RRAM-based non-volatile associative memory that naturally enables long-term feature embedding and efficient feature retrieval. In our MANN system, the frontend NN is initially trained offline (*meta-training*), after which its weights are fixed and do not need to be updated [1]. During one-shot learning, novel features (from unseen classes that are not included in NN meta-training) are mapped into the associative memory, using only one example per class. During inference, the associative memory performs similarity-based retrieval given query samples and makes predictions based on similarity.

Fully Integrated RRAM-CMOS Associative Memory

Memory cell-level explorations towards new hardware architectures will not be able to answer key questions regarding real-time chip behaviors for target applications [2]. Here, we present a 64-kbit fully-integrated RRAM-CMOS associative memory (AM) chip as the backend of MANNs. The AM core occupies 0.2 mm^2 at 40 nm technology node [3]. Our chip supports the key one-shot learning and inference operations needed for a MANN through two modes: (1) feature vector embedding within AM; and (2) L1 distance computation between query set (test images) and support set (embedded novel features) for similarity-based prediction. Fig. 2 illustrates the overall chip architecture for the 64-kbit AM core. Feature embedding mode is enabled by operating the AM core as a typical 64-kbit random access memory. The 2T-2R memory cells encode data in a complementary fashion [4], i.e., high resistance state (HRS)-low resistance state (LRS) encodes bit '0' while LRS-HRS encodes bit '1'. As a result, 128 bits per feature vector can be stored along bitline (BL) direction. The AM core is partitioned into 8 sub-AM banks that can be used independently. Every set of 8 rows shares a sense amplifier (SA), and partial L1 distance results captured by sensing circuitry are accumulated for similarity-based prediction. Fig. 3 shows the RRAM array bias schemes used in programming and sensing of vectors. With write-verify programming applied across full 64-kbit chip, array-level resistance distributions (50% cells programmed to HRS and 50% to LRS) are obtained and shown in Fig. 4. We choose to trade off HRS uniformity for larger memory window (HRS above 100 kΩ) while keeping LRS in a tight distribution. This is relevant to the sensing circuitry during L1 distance computation. We implement the approximate search using two simple inverter-based sense amplifiers (SAs) with different Vth to support sensing 3 levels of voltage (Fig. 5). Within the context of similarity measurement using memory circuits, our chip achieves an energy consumption of 270 pJ for searching among 32 128-bit vectors. Fig. 6 shows how this compares with other reported memory chips that only support exact search [6]-[10], using the same workload of 32-entry 128-bit approximate search supported natively by our AM core. Compact cell structure reduces total wire length and thus dynamic energy when multiple rows or columns are activated in parallel.

One-Shot Learning Experiments on AM Chip

We demonstrate one-shot learning and inference on a widely used Omniglot dataset [11], which consists of 1623 characters from 50 alphabets, each drawn by 20 different persons. The learning task on Omniglot reflects human-learning scenarios with very few examples per class, and had never been demonstrated on hardware chips before. In our chip demonstrations, 32-way 1-shot learning is chosen as a much harder task than the commonly used 20-way or 10-way tasks [1], [11]. Fig. 7 shows the flow of meta-training, one-shot learning and inference. We implement the frontend 4-layer NN in software for feature extraction, which is pre-trained (meta-training phase) with full precision and its weights are fixed. Extracted feature vectors are then quantized and mapped onto the AM subarrays as 4-bit thermometer codes. One 128-bit vector is embedded along one BL (256 RRAM cells), corresponding to one unique feature vector. For one-shot learning demonstration on chip, we pick 32 unseen classes (not present in frontend meta-training) from 212 classes. Only 1 image from each of the 32 classes is used for learning. Fig. 8 shows the complete map of learned feature vectors in the 64-kbit AM core, encoded by the 2T-2R RRAM resistance distributions. Running Omniglot inference tests, our chip achieves a measured ~72% accuracy (Fig. 9). Note that the 32-way, 1-shot task is a more difficult task than the commonly reported 20-way or 10-way learning on Omniglot on software [11]. We test the robustness of chip inferencing by continuously running >6 million examples (cycling 320 test images), and monitoring the prediction accuracies from chip measurements (Fig. 10). Fig. 11 shows the measured power-frequency scaling of the AM core. Taking in-memory L1 distance computation and prediction as the basic operations, the chip reaches 118 GOPS/W energy efficiency. Fig. 12 and Fig. 13 show the measurement setup and the chip micrograph. Finally, Table I summarizes the key characteristics of our chip.

Conclusion

We report a 64-kbit, 118 GOPS/W non-volatile associative memory chip that demonstrates on-chip one-shot learning with ~72% measured hardware accuracy on the challenging 32-way 1-shot learning task on Omniglot benchmark. This work, the first hardware chip demonstration of one-shot learning, leads towards future energy-efficient hardware learning machines with continuous, lifelong learning capabilities.

Acknowledgements We gratefully acknowledge chip implementation support and collaboration with W.S. Khwa, M.F. Chang, Y.D. Chih, H. Chuang and their colleagues at TSMC. This work is supported in part by SRC JUMP ASCENT Center and Stanford SystemX Alliance. A. Levy is supported by NSF GRFP. We also thank G. Lallement, R. Radway, K. Prabhu, Prof. S. Mitra and Prof. B. Murmann (Stanford University).

References: [1] O. Vinyals *et al.*, *NeurIPS*, 2016. [2] K. Ni *et al.*, *Nature Electronics*, 2019. [3] C.-C. Chou, *et al.*, *ISSCC*, 2018. [4] J. Li, *et al.*, *VLSI Tech.*, 2013. [6] M. Yabuuchi, *et al.*, *VLSI Circuits*, 2018. [7] C.-X. Yue, *et al.*, *ASSCC*, 2018. [8] S. Jeloka, *et al.*, *VLSI Circuits*, 2015. [9] C. C. Lin, *et al.*, *ISSCC*, 2016. [10] M. F. Chang, *et al.*, *ISSCC*, 2015. [11] B.M. Lake, *Science*, 2015.

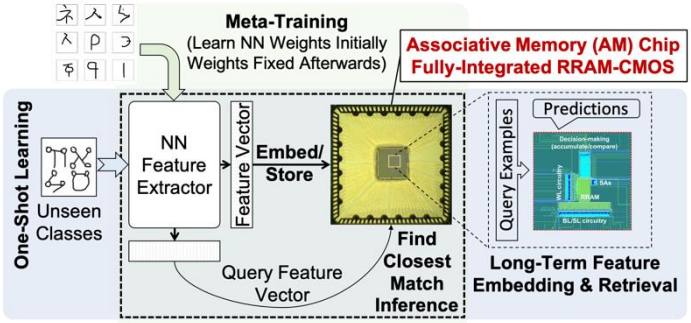

Fig. 1. Overview of our MANN system for one-shot learning. Front-end NN is off-chip and only requires one-time offline training (*meta-training phase*). Extracted features from unseen classes are embedded into the associative memory (*one-shot learning phase)*, while additional query examples are used for retrieval/prediction of labels (*inference phase*).

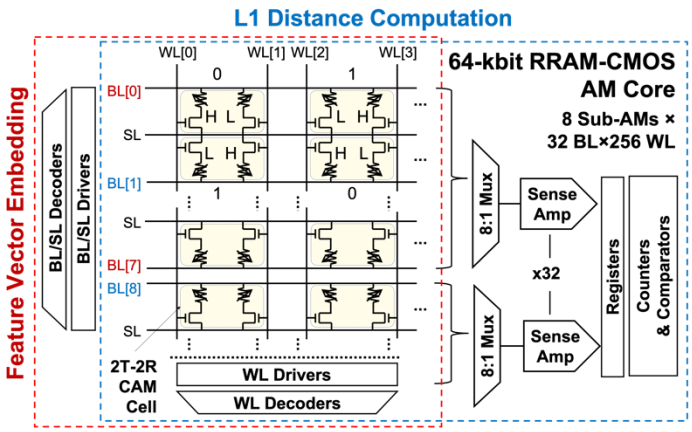

Fig. 2. 64-kbit RRAM-based non-volatile AM. Feature embedding and L1 distance computation enable one-shot learning and inference.

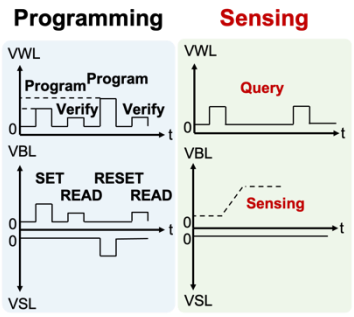

Fig. 3. RRAM biasing scheme during programming and sensing. SET-RESET write-verify is used for programming. Sensing uses WLs to send in query data while BLs are driven by SA circuitry.

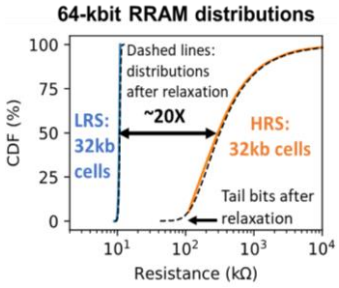

Fig. 4. Measured array-level low and high resistance (LRS, HRS) distributions from 64-kbit RRAMs. The overlaid dashed lines indicate the measured resistances from the same array after relaxation.

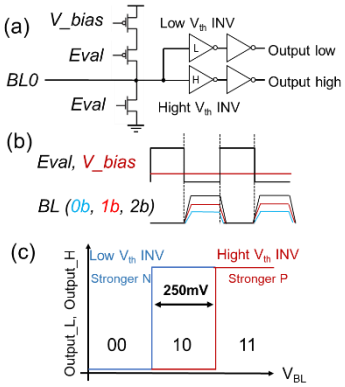

Fig. 5. Sense amplifier (SA) circuit and sensing mechanism. V_bias and SA supply voltage are tuned for each operating freq.

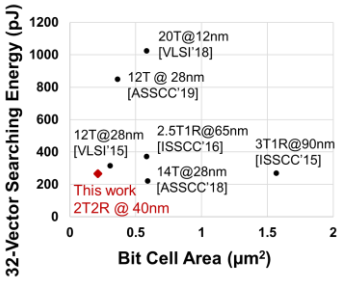

Fig. 6. 32-vector searching energy and bit cell area for various memory chips with searching capabilities, using the same 32-entry 128-bit approximate search workload.

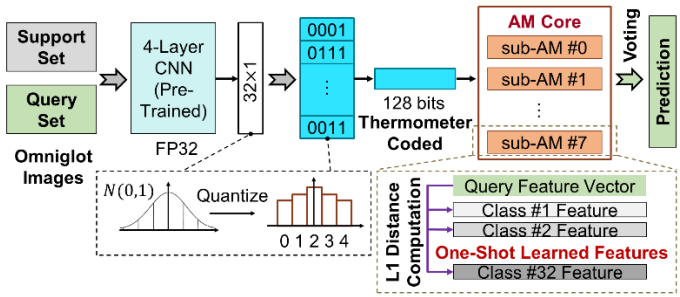

Fig. 7. Flow of meta-training, one-shot learning and inference. Pre-trained 4-layer CNN extracts feature vectors that are quantized and encoded into 4-bit thermometer codes. During one-shot learning, novel feature vectors (from support set) are embedded into all sub-AM banks. During inference, L1 distance computation is performed given query vectors for similarity-based prediction, while utilizing all the sub-AM banks for final voting (ensembling, off-chip).

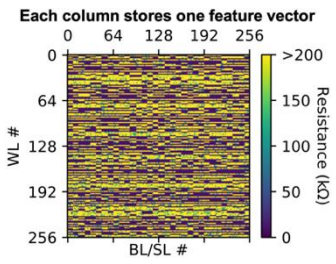

Fig. 8. 64-kbit data pattern (resistance distributions) after 32-way, 1-shot learning on chip. 32 novel features are broadcast to 8 subarrays and programmed as vectors along the BLs, for a total of 256 feature vectors.

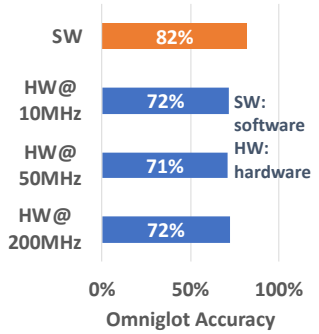

Fig. 9. Measured inference accuracy across different chip frequencies tested for 32-way, 1-shot learning task on the Omniglot benchmark.

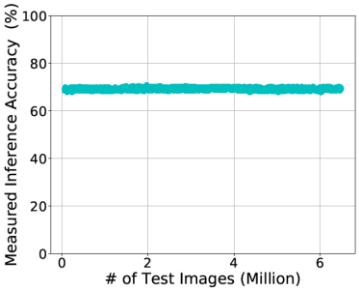

Fig. 10. Measured chip inference accuracy on a continuous stream of > 6 million images (cycling 320 Omniglot test images), demonstrating chip robustness.

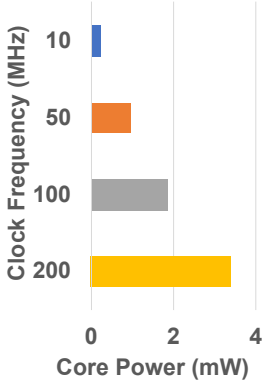

Fig. 11. Power-freq. scaling of the AM core up to 200 MHz.

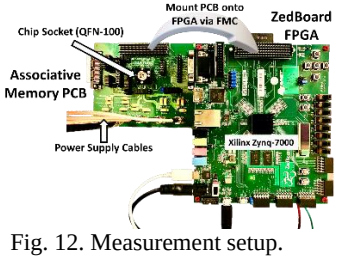

Fig. 12. Measurement setup.

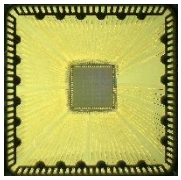

Fig. 13. Chip micrograph.

Table I. Chip summary.

RRAM-Based AM Chip for One-Shot Learning and Inference	
Technology Node	40 nm
Capacity	64 kbits
Core Area	0.2 mm²
Max. Frequency	200 MHz
Energy Efficiency (L1 dist. & predict)	118 GOPS/W
Accuracy on Omniglot (32-way, 1-shot)	72%