30.10 A 1TOPS/W Analog Deep Machine-Learning Engine with Floating-Gate Storage in 0.13μm CMOS

Junjie Lu, Steven Young, Itamar Arel, Jeremy Holleman

University of Tennessee, Knoxville, TN

Direct processing of raw high-dimensional data such as images and video by machine learning systems is impractical both due to prohibitive power consumption and the "curse of dimensionality," which makes learning tasks exponentially more difficult as dimension increases. Deep machine learning (DML) mimics the hierarchical presentation of information in the human brain to achieve robust automated feature extraction, reducing the dimension of such data. However, the computational complexity of DML systems limits large-scale implementations in standard digital computers. Custom analog or mixed-mode signal processors have been reported to yield much higher energy efficiency than DSP [1-4], presenting the means of overcoming these limitations. However, the use of volatile digital memory in [1-3] precludes their use in intermittently-powered devices, and the required interfacing and internal A/D/A conversions add power and area overhead. Nonvolatile storage is employed in [4], but the lack of learning capability requires task-specific programming before operation, and precludes online adaptation.

The feasibility of analog clustering, a key component of DML, has been demonstrated in [5]. In this paper, we present an analog DML engine (ADE) implementing DeSTIN [6], a state-of-art DML framework, and featuring online unsupervised trainability. Floating-gate nonvolatile memory facilitates operation with intermittent harvested energy. An energy efficiency of 1TOPS/W is achieved through massive parallelism, deep weak-inversion biasing, current-mode analog arithmetic, distributed memory, and power gating applied to per-operation partitions. Additionally, algorithm-level feedback desensitizes the system to errors such as offset and noise, allowing reduced device sizes and bias currents.

Figure 30.10.1 shows the architecture of the ADE, in which seven identical cortical circuits (nodes) form a 4-2-1 hierarchy. Each node captures regularities in its inputs through an unsupervised learning process. The lowest layer receives raw data (e.g. the pixels of an image), and continuously constructs belief states that characterize the sequence observed. The inputs of nodes on 2nd and 3rd layers are the belief states of nodes at their respective lower layers. The beliefs of the top layer are then used as rich features for a classifier.

The node (Fig. 30.10.2) incorporates an 8×4 array of reconfigurable analog computation cells (RAC), grouped into 4 centroids, each with 8-dimensional input. The centroids are characterized by their mean μ and variance σ^2 in each dimension, stored in their respective floating gate memories (FGM). In a training cycle, the analog arithmetic elements (AAE) calculate a simplified Mahalanobis distance (assuming a diagonal covariance matrix) D_{MAH} between the input observation $\boldsymbol{o}$ and each centroid. The 8-D distances are built by joining the output currents. A distance processing unit (DPU) performs inverse-normalization (IN) operation to the 4 distances to construct the belief states, which are the likelihood that the input belongs to each centroid. Then the centroid parameters μ and σ^2 are adapted using the online clustering algorithm. The centroid with the smallest Euclidean distance D_{EUC} to the input is selected (classification). The errors between the selected centroids and input are loaded to the training control (TC) and their memories are then updated proportionally. In recognition mode, only the belief states are constructed and the memories are not adapted. Intra-cycle power gating is applied to reduce the power consumption by up to 37%.

Figure 30.10.3 shows the schematic of the RAC, which performs three different operations through reconfigurable current routing. Two embedded FGMs provide nonvolatile storage for centroid parameters. Capacitive feedback stabilizes the floating gate voltage (V_{FG}) to yield pulse-width controlled update. Tunneling is enabled by lowering its supply to bring down the V_{FG}, increasing the voltage across the tunneling junction. Injection is enabled by raising the source of the injection transistor. This allows random-accessible bidirectional updates without the need for on-chip high-voltage switches or charge pump. A 2-T V-I converter then provides a current output and sigmoid update rule. The FGM consumes 0.5nA of bias current, and shows an 8b programming accuracy and a 46dB SNR at full scale. The absolute value circuit (ABS) in the AAE rectifies the bidirectional difference current between o and μ. Class-B operation and the virtual ground provided by amplifier A allow high-speed resolution of small current differences. The rectified currents are then fed into a translinear X^2/Y circuit, which simulations indicate operates with more than an order of magnitude higher energy efficiency than its digital equivalence.

In the belief construction phase, the DPU (Fig. 30.10.4) inverts the distance outputs from the 4 centroids to calculate similarities, and normalizes them to yield a valid probability distribution. The output belief states are sampled then held for the rest of the cycle to allow parallel operation of all layers. The sampling switch reduces current sampling error due to charge injection: a diode-connected PMOS provides a reduced V_{GS} to the switch NMOS to turn it on with minimal channel charge. The S/H achieved less than 0.7mV of charge injection error (2% current error), and less than 14μV of droop with parasitic capacitors as holding capacitor. In classification phase, the IN circuits are reused together with the winner-take-all network (WTA) to classify the observation to the nearest centroid. A starvation trace (ST) circuit is implemented to address unfavorable initial conditions wherein some centroids are starved of nearby inputs and never updated. The ST provides starved centroids with a small but increasing additional current to force their occasional selection and pull them into more populated areas of the input space. The lower right of Fig. 30.10.4 shows the TC circuit, which performs current-to-pulse-width conversion using a V_{DD}-referenced comparison. Proportional updates cause the mean and variance memories to converge to the sample statistics, respectively.

The ADE is evaluated on a custom test board with data acquisition hardware connecting to a host PC. The waveforms in Fig. 30.10.5 show the measured beliefs, one from each layer. The sampling of beliefs proceeds from the top layer to the bottom to avoid delays due to output settling. The performance of the node is demonstrated by solving a clustering problem. The input data consists of 4 underlying clusters, each drawn from a Gaussian distribution with different mean and variance. The node achieves accurate extraction of the cluster parameters (μ and σ^2), and the ST ensures a robust operation against unfavorable initial conditions.

We demonstrate feature extraction for pattern recognition with the setup shown in Fig. 30.10.6. The input patterns are 16×16 symbol bitmaps corrupted by random pixel errors. An 8×4 moving window defines the pixels applied to the ADE's 32-D input. First the ADE is trained unsupervised with examples of patterns. After the training converges, the 4 belief states from the top layer are used as rich features and classified with a neural network implemented in software, achieving a dimension reduction from 32 to 4. Recognition accuracies of 100% with corruption lower than 10%, and 95.4% with 20% corruption are obtained, comparable to a software baseline, demonstrating robustness to the non-idealities of analog computation.

The ADE was fabricated in a 0.13μm CMOS process with thick-oxide IO FETs. The die micrograph is shown in Fig. 30.10.7, together with a performance summary and a comparison with state-of-art bio-inspired parallel processors utilizing analog computation. We achieve 1TOPS/W peak energy efficiency in recognition mode. Compared to state-of-art, this work achieves very high energy efficiency in both modes. This combined with the advantages of nonvolatile memory and unsupervised online trainability makes it a general-purpose feature extraction engine ideal for autonomous sensory applications or as a building block for large-scale learning systems.

References:

[1] J. Park, et al., "A 646GOPS/W Multi-Classifier Many-Core Processor with Cortex-Like Architecture for Super-Resolution Recognition," *ISSCC Dig. Tech. Papers,* pp. 168-169, Feb. 2013.

[2] J. Oh, et al., "A 57mW Embedded Mixed-Mode Neuro-Fuzzy Accelerator for Intelligent Multi-Core Processor," *ISSCC Dig. Tech. Papers,* pp. 130-132, Feb. 2011.

[3] J.-Y. Kim, et al., "A 201.4GOPS 496mW Real-Time Multi-Object Recognition Processor With Bio-Inspired Neural Perception Engine," *ISSCC Dig. Tech. Papers,* pp. 150-151, Feb. 2009.

[4] S. Chakrabartty and G. Cauwenberghs, "Sub-Microwatt Analog VLSI Trainable Pattern Classifier," *IEEE J. Solid-State Circuits,* vol. 42, no. 5, pp. 1169-1179, May 2007.

[5] J. Lu, et al., "An Analog Online Clustering Circuit in 130nm CMOS," *IEEE Asian Solid-State Circuits Conference,* Nov. 2013.

[6] S. Young, et al., "Hierarchical Spatiotemporal Feature Extraction using Recurrent Online Clustering," Pattern Recognition Letters, Oct. 2013.

 978-1-4799-0920-9/14/$31.00

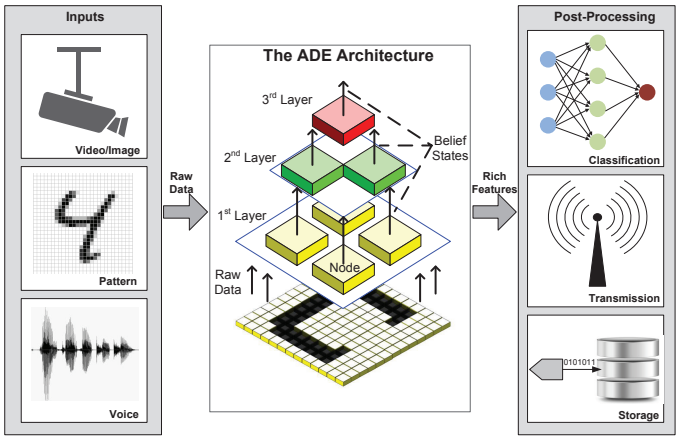


Figure 30.10.1: The analog deep learning engine (ADE) architecture.

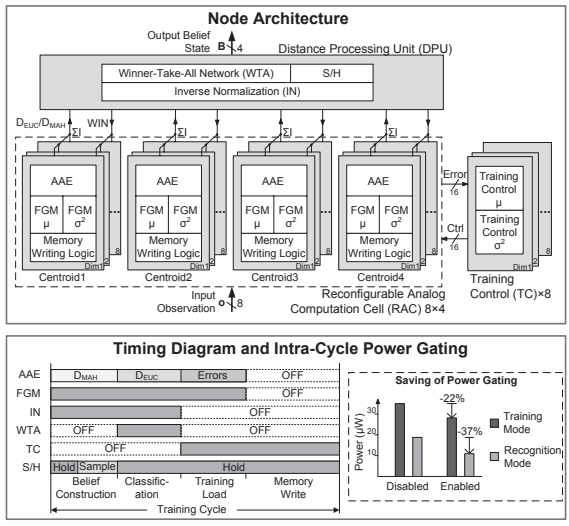


Figure 30.10.2: The node architecture and its timing diagram showing power gating.

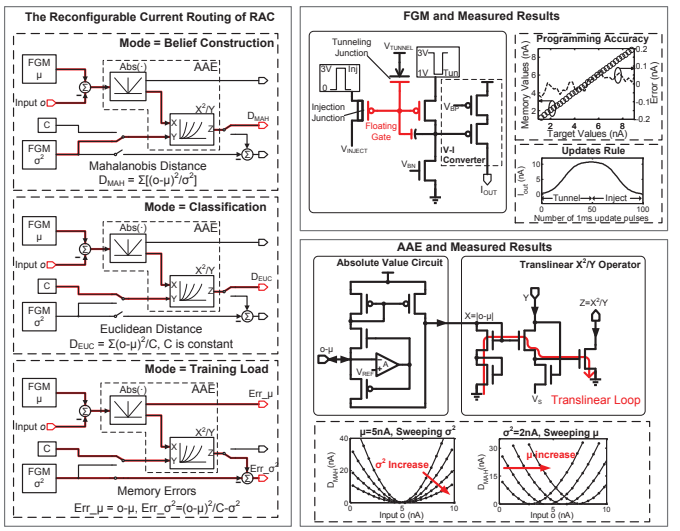


Figure 30.10.3: The reconfigurable current routing of the RAC, the schematics of the FGM and AAE and measurement results.

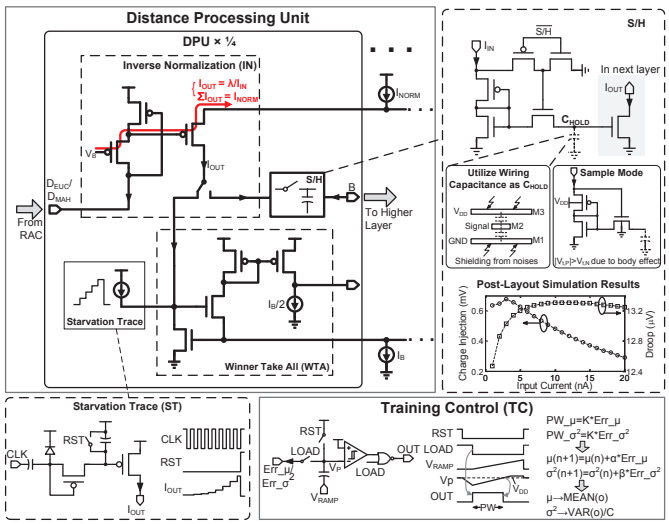


Figure 30.10.4: The schematic of the DPU and its sub-blocks. The training control is shown on the lower right.

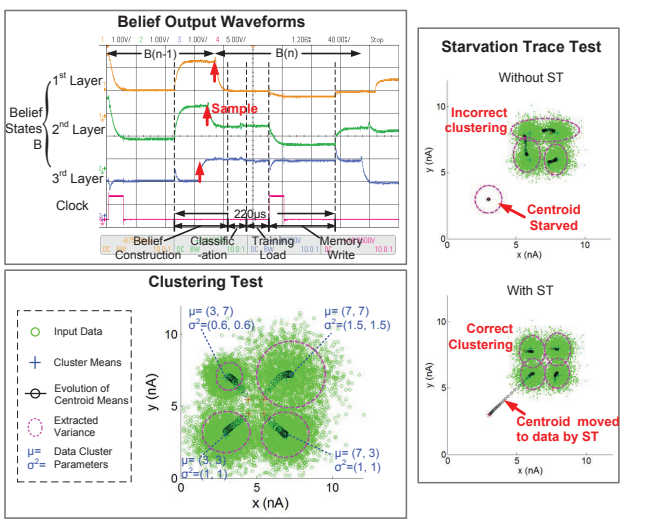


Figure 30.10.5: Measured output waveforms, clustering and ST test results. For clustering, 2-D results are shown for better visualization.

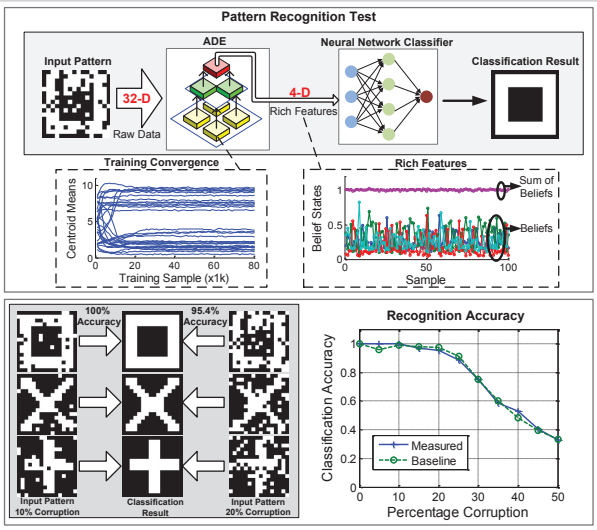
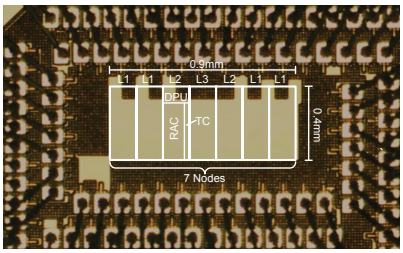


Figure 30.10.6: Pattern recognition test setup and results, demonstrating accuracy comparable to baseline software simulation.



Technology	1P8M 0.13μm CMOS	
Power Supply	3V	
Active Area	0.9mm×0.4mm	
Memory	Non-Volatile Floating Gate	
Memory SNR	46dB	
Training Algorithm	Unsupervised Online Clustering	
Output Feature	Inverse-Nomalized Mahanalobis Distances	
Input Referred Noise	56.23pA$_{rms}$	
System SNR	45dB	
I/O Type	Analog Current	
Operating Frequency	Training Mode	4.5kHz
	Recognition Mode	8.3kHz
Power Consumption	Training Mode	27μW
	Recognition Mode	11.4μW
Energy Efficiency	Training Mode	480GOPS/W
	Recognition Mode	1.04TOPS/W

	This work	ISSCC'13 [1]	ISSCC'11 [2]	ISSCC'09 [3]
Process	0.13μm	0.13μm	0.13μm	0.13μm
Purpose	DML Feature Extraction	Object Recognition	Neural-Fuzzy Accelarator	Object Recognition
Non-volatile Memory	Floating Gate	NA	NA	NA
Power (W)	11.4μW	260mW	57mW	496mW
Peak Energy Efficiency	1.04TOPS/W	646GOPS/W	655GOPS/W	290GOPS/W

Figure 30.10.7: Die micrograph, performance summary and comparison table.

978-1-4799-0920-9/14/$31.00