A 6.67mW Sparse Coding ASIC Enabling On-Chip Learning and Inference

Jung Kuk Kim, Phil Knag, Thomas Chen, Zhengya Zhang
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor

Abstract

A sparse coding ASIC is designed to learn visual receptive fields and infer the sparse representation of images for encoding, feature detection and recognition. 256 leaky integrate-and-fire neurons are connected in a 2-layer network of 2D local grids linked in a 4-stage systolic ring to reduce the communication latency. Spike collisions are kept sparse enough to be tolerated to save power. Memory is divided into a core section to support inference, and an auxiliary section that is only powered on for learning. An approximate learning tracks only significant neuron activities to save memory and power. The 3.06mm^2 65nm CMOS ASIC achieves an inference throughput of 1.24Gpixel/s at 1.0V and 310MHz, and on-chip learning can be completed in seconds. Memory supply voltage can be reduced to 440mV to exploit the soft algorithm that tolerates errors, reducing the inference power to 6.67mW for a 140Mpixel/s throughput at 35MHz.

Introduction

Visual receptive field (RF) of a neuron [1] is a region of space in which the presence of a stimulus alters the firing of the neuron. RFs can be understood as features or basis functions of images. Sparse and independent local network (SAILnet) [2] is a machine learning algorithm known as sparse coding [1] that learns RFs through training a network of model neurons, and infers the sparse representation of the input image using the most salient RFs (Fig. 1). Inference based on the learned RFs enables efficient image encoding, and detecting features and objects [3], [4]. However, the implementation of an energy-efficient high-throughput sparse coding processor faces challenges of on-chip interconnect and memory bandwidth to support the parallel operations of hundreds or more model neurons. Existing hardware designs cannot be adapted for sparse coding [5], [6], and they often resort to off-chip memory and processing [6]-[8].

Two-Layer Grid-Ring Network Architecture

We develop the first fully integrated sparse coding ASIC that consists of 256 digital neurons, 64K feed-forward synapses, and 64K feedback synapses. The sparse coding chip performs both unsupervised learning and inference on-chip. The RF of each model neuron is initialized with random noise, and each neuron learns its RF through training images. After learning converges, the chip is able to perform inference to encode images by the sparse activation of neurons, i.e., neuron spikes. To check the fidelity of inference, the input image can be compared with its reconstruction by the weighted sum of the RFs of the activated neurons (Fig. 2).

Each model neuron in the sparse coding chip is a compute node that performs leaky integrate-and-fire [2]. A two-layer network is designed to allow all neurons to communicate efficiently: a cluster of 64 neurons are connected in a 2D grid (Fig. 3) that improves the communication delay over a 1D bus; and the root nodes of four grids are connected in a 4-stage systolic ring (Fig. 4). The grid size is designed to limit the wire loading for sub-ns timing and bound the spike collision rate; and the ring is kept short to reduce latency. We exploit the soft algorithm to tolerate occasional spike collisions. Collisions are detected and tolerated to save power, and we verify that a 5% or lower collision rate is tolerated without causing any noticeable degradation in fidelity.

Memory Partition and Approximate Learning

Each model neuron stores RFs, termed Q weights (256 entries for a 16×16 image patch), and synaptic strengths, termed W weights (256 entries, one for each neuron). The Q and W weights of a cluster of 64 neurons are stored together in Q and W memory for a higher efficiency. We optimize the word length of Q and W weight to 11 and 8 bits, respectively, to minimize storage and guarantee reliable convergence. After learning converges, the Q and W weight can be further shortened to 4 bits each for inference with minimal loss in fidelity. We exploit the difference in word length requirements between learning and inference to partition the Q and W memory to two sections that are placed on separate supply rails: the core section to support inference, and the auxiliary section that is only powered on for learning. The core memory is implemented in high-bandwidth register file to support real-time inference. The auxiliary memory is implemented in a lower-bandwidth SRAM to provide the extra bits needed for learning. As learning is called less frequently, the SRAM power consumption becomes negligible. On-chip learning is nonetheless orders of magnitude faster and lower power than off-chip learning.

Learning and inference make use of the same network of model neurons, but learning also updates Q and W weights, which dictates the learning speed. Learning is done by a snooping core attached to the top-level ring to listen to neuron spikes and record the activities in a cache (Fig. 4). After a batch of training images, the snooping core reads the cache and makes weight adjustments according the SAILnet learning rules [2]. To accelerate learning and reduce the cache size, we implement approximate learning to record the activities of only the first 10 neurons that spike for each input image patch. Experimental evidence points to the fact that the neurons that spike first tend to be the most active. The remaining neuron activities play a minor role, and can be safely ignored.

Chip Measurement and Error Tolerance

The sparse coding ASIC test chip is implemented in TSMC 65nm CMOS (Fig. 7). Input images are scanned in to an SRAM for testing. Inference operates at a maximum 310MHz, consuming 218mW at 1.0V and room temperature. Inference is carried out in steps for each 16×16 input image patch. For a high fidelity, the number of steps is set to at least 64, which translates to a maximum inference throughput of 1.24Gpixel/s (Gpx/s) at an energy efficiency of 176pJ/px (Fig. 5, Table I). To enable learning, the auxiliary memory is powered on. Learning consumes 228mW at 1.0V and 235MHz for a learning speed of 188Mpx/s (Fig. 6, Table I). A training set of 1 million 16×16 image patches is completed in 1.4s.

The sparse coding algorithm is error tolerant, and with on-chip learning, errors can be corrected by on-line training. Our measurements indicate a gradual degradation of the normalized root-mean-square error (NRMSE) of the reconstructed image (measure of fidelity) until the core memory supply is lowered to 390mV, where a nearly 10^{-3} core memory bit error rate results in no more than 0.03 NRMSE in inference (Fig. 8). The error tolerance is exploited to reduce power. The core memory supply voltage can be reduced to 440mV, while still keeping NRMSE within 0.01. Together with voltage scaling the core logic, the inference power consumption is reduced to 6.67mW for an inference throughput of 140Mpx/s, improving the energy efficiency 48pJ/px (Fig. 5, Table I). Learning requires writing to memory, which places lower bounds on the core and auxiliary memory supply at 580mV and 600mV, respectively. At these low supplies, the learning power consumption is reduced to 6.8mW for a learning speed of 16Mpx/s (Fig. 6, Table I). A comparison with recent literature is presented in Table II. The on-chip learning capability, as well as the achieved high throughput and energy efficiency demonstrate the potential of the sparse coding ASIC for embedded vision processing tasks.

Acknowledgements

The work is supported in part by DARPA. We thank Wei Lu, Michael Flynn, and Garrett Kenyon for helpful suggestions.

References

[1] Olshausen and Field, *Nature*, vol. 381, no. 6583, pp. 607-609, 1996.
[2] Zylberberg *et al.*, *PLoS Computational Biology*, vol. 7, no. 10, pp. 1-12, 2011.
[3] Chalupa and Werner, Eds., *The Visual Neurosciences*, 2003.
[4] Field, *Neural Computation*, vol. 6, no. 4, pp. 559-601, 1994.
[5] Seo *et al*., *CICC*, 2011.
[6] Merolla *et al*., *CICC*, 2011.
[7] Vogelstein *et al*., *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 253-265, 2007.
[8] Choudhary *et al*., *ICANN*, 2012, pp. 121-128.

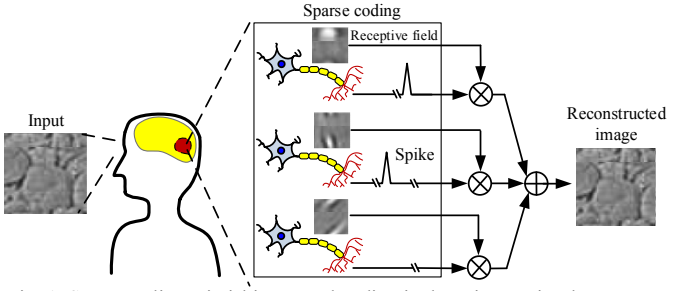

Fig. 1. Sparse coding mimicking neural coding in the primary visual cortex. The input image can be reconstructed by the weighted sum of receptive fields of model neurons.

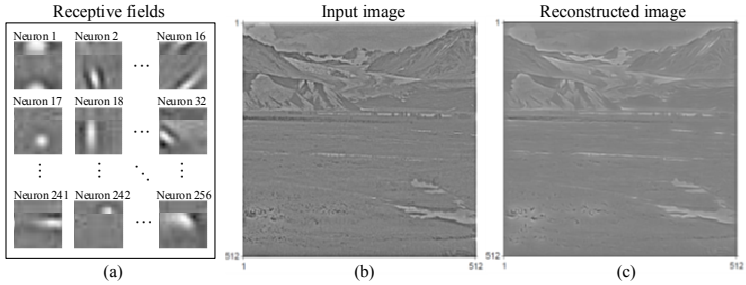

Fig. 2. (a) Receptive fields learned by model neurons through training, (b) an input image presented to the sparse coding ASIC, and (c) the reconstructed image based on the neuron spikes obtained by inference.

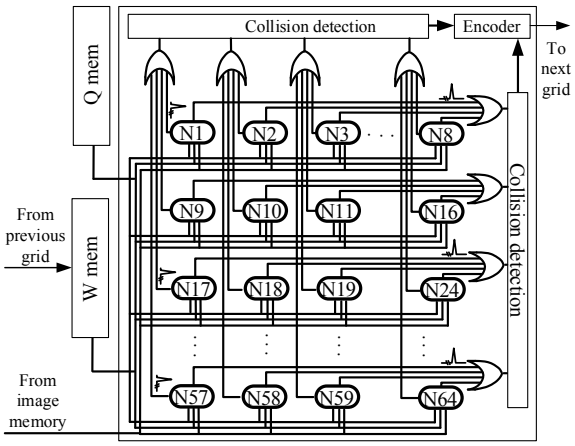

Fig. 3. 2D grid of a cluster of 64 neurons. Spike collisions are detected and tolerated to save power.

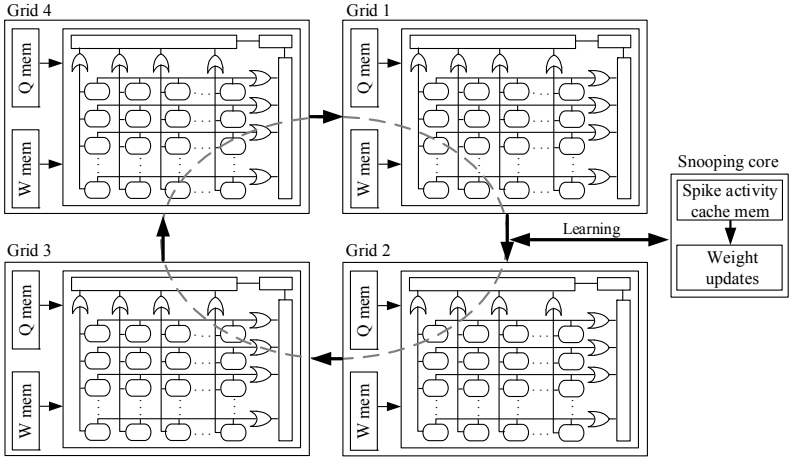

Fig. 4. 4-stage systolic ring connecting 4 2D local grids. A snooping core is attached to the ring to record neuron spikes for learning.

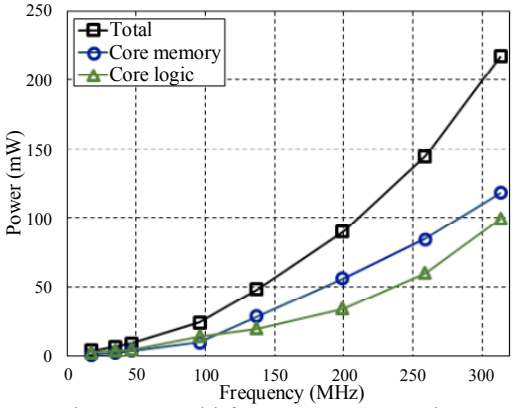

Fig. 5. Measured inference power consumption: core memory power, core logic power, and total inference power (auxiliary memory is powered off in inference). Power is measured at the minimum logic and memory supply voltages for each clock frequency.

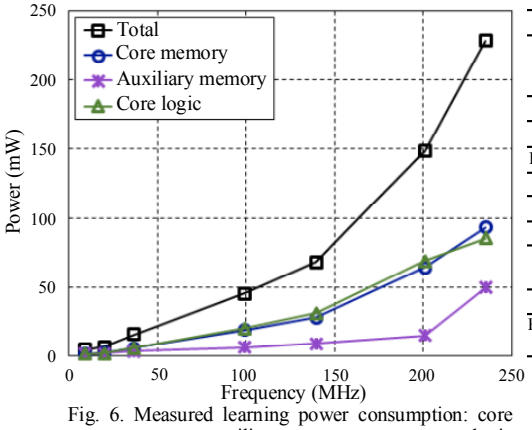

Fig. 6. Measured learning power consumption: core memory power, auxiliary memory power, core logic power, and total learning power (auxiliary memory is powered on in learning). Power is measured at the minimum logic and memory supply voltages for each clock frequency.

TABLE I: CHIP SUMMARY

Technology	TSMC 65nm GP CMOS			
Core Area	1.75mm × 1.75mm (Core logic: 1.16mm^2, Core mem: 1.01mm^2, Aux. mem: 0.89mm^2)			
Chip Area	2.11 × 2.11mm (4.45mm^2)			
	Inference		Learning	
Frequency (MHz)	35	310	20	235
Core logic (V)	0.53	1.00	0.50	1.00
Core mem (V)	0.44	1.00	0.58	1.00
Aux. mem (V)	0.00	0.00	0.60	1.00
Throughput (Mpixel/s)	140	1240	16	188
Power (mW)	6.67	218	6.83	228.1
Energy Efficiency (pJ/pixel)	47.6	175.8	426.9	1213

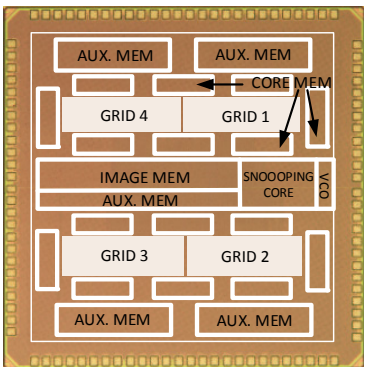

Fig. 7. Chip microphotograph

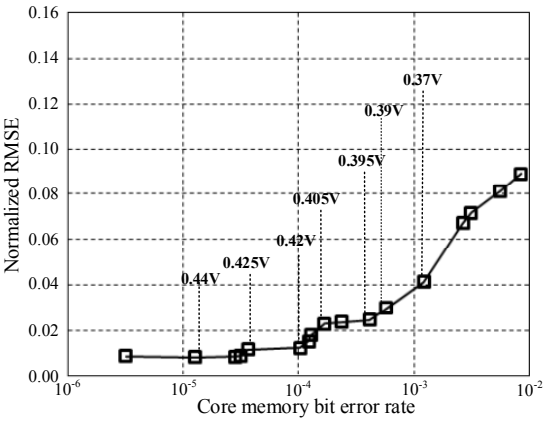

Fig. 8. Measured normalized root-mean-square error (NRMSE) in inference with increasing core memory bit error rate. The core memory supply voltage is annotated.

TABLE II: COMPARISION WITH PRIOR WORKS

Reference	Seo [5]	Merolla [6]	This work
# Neurons	256	256	256
# Synapses	64K	256K	128K
Bitwith of a Synapse	4 bits	1 bit	8 and 11 bits
Mem size	256Kbits	256Kbits	1.31Mbits
Interconnect	Crossbar	Crossbar	2-layer grid and ring
Algorithm	STDP	RBM	SAILnet
Learning	On chip	Off chip	On chip
Application	Pattern recognition	Digit recognition	Image sparse coding
Technology	IBM 45nm	IBM 45nm	TSMC 65nm
Core Area	4.2mm^2	4.2mm^2	3.1mm^2
Energy metric	-	45pJ/spike	48pJ/pixel