A 640M pixel/s 3.65mW Sparse Event-Driven Neuromorphic Object Recognition Processor with On-Chip Learning

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

Abstract

A 1.82mm^2 65nm neuromorphic object recognition processor is designed using a sparse feature extraction inference module (IM) and a task-driven dictionary classifier. To achieve a high throughput, the 256-neuron IM is organized in four parallel neural networks to process four image patches and generate sparse neuron spikes. The on-chip classifier is activated by sparse neuron spikes to infer the object class, reducing its power by 88% and simplifying its implementation by removing all multiplications. A light-weight co-processor performs efficient on-chip learning by taking advantage of sparse neuron activity to save 84% of its workload and power. The test chip processes 10.16G pixel/s, dissipating 268mW. Integrated IM and classifier provides extra error tolerance for voltage scaling, lowering power to 3.65mW at a throughput of 640M pixel/s.

Introduction

Recognizing objects in an image can be accomplished by first extracting features from the image using an inference module (IM), and then classifying the object based on the extracted features using a classifier (Fig. 1). The locally competitive algorithm (LCA) [1] is a neural-inspired IM that infers a sparse set of features to best represent the input image. Compared to a conventional feature extraction algorithm, e.g., SIFT [2], an LCA-based IM simplifies the classifier and potentially improves the classification accuracy. Past work has produced an 18-neuron spiking LCA based analog IM [3] (Fig. 2), but the small scale is not suitable for practical problems. A 256-neuron digital IM using SAILnet [4] is scalable and achieved a much higher throughput (Fig. 2), but the design was dominated by memory and it is not capable of object classification. In this work, we demonstrate a 256-neuron 10.16G pixel/s spiking LCA based IM that is integrated with a task-driven dictionary classifier to exploit the sparse feature extraction for an end-to-end object recognition processor. A light-weight learning co-processor is integrated on chip to learn and adapt the feature library. This self-contained design empowers low-power and high-performance embedded applications of computer vision.

Sparse Feature Extraction and Sparse Event-Driven Classifier

The 256-neuron IM is organized in four 64-neuron spiking neural networks, each operating on a 16×16 input image patch to extract features in parallel (Fig. 3). Through training, each neuron develops a 16×16 receptive field (RF), i.e., feature that excites the neuron. The division of an input image into smaller patches reduces the size of the neural networks, resulting in a 4× reduction in memory size. By deploying the four neural networks in parallel, a high throughput and comparable inference accuracy can be achieved without the memory overhead associated with a large neural network. The neuron dynamics are tuned to achieve a high inference accuracy with less than 16% of neurons firing over a 2τ inference period (τ: neuron time constant), saving the power and enabling a sparse event-driven classifier and a light-weight on-chip learning.

Each IM network encompasses 64 digital integrate-and-fire neurons. Every 8 neurons are connected in a grid for detecting neuron spikes and generating address events (AE) to signal neuron spikes. The 8 grids are connected in an 8-stage systolic ring to propagate AEs (Fig. 3). The network structure represents the optimal tradeoff between hardware efficiency and inference accuracy: a large grid is more compact, but results in more simultaneous neuron spikes colliding over the grid, worsening the inference accuracy; while a systolic ring preserves neuron spikes but a long ring costs more area and power.

A real-time classifier implementing task-driven dictionary learning [5] is tightly integrated with the IM to recognize objects from 10 classes. Four sub-classifiers are each attached to an IM network by tapping the systolic ring (Fig. 3). Each sub-classifier consists of 10 class nodes listening to the neuron spikes generated by an IM network. A neuron spike represents an active feature that triggers a weighted vote for each class node. The weight depends on the degree of the feature's association with the object class, and they are learned through supervised learning. Since neuron spikes are sparse, the classifier is designed to be event-driven to reduce its power by 88%. The binary spike train allows the classifier to be implemented with adders, replacing costly multipliers to save 72% area and 65% power. The class node outputs from the four sub-classifiers are used to score the most likely object class as output.

Light-Weight On-Chip Learning Co-Processor

Real-time learning is not necessary for practical applications, but on-chip learning reduces I/O power and it provides quick adaptation to changing environment. For this reason, a light-weight learning co-processor is integrated on chip to implement the rules governing the learning of the 64 RFs that form the feature dictionary of the IM, and 16K feedback weights between neurons. The RFs are developed iteratively following stochastic gradient descent to minimize image encoding error and improve sparsity.

Learning involves large vector and matrix multiplications that are naturally mapped to a vector processor. However, the vectors are sparse due to sparse neuron spikes (Fig. 4). We take advantage of this insight to design a scalar processor to cut over 84% of the workload and power. The low-cost scalar learning co-processor provides three instructions to support learning: vector-matrix product, matrix scaling, and matrix-matrix product, which are all executed element-by-element in a serial fashion.

Test Chip Measurement

A test chip of the object recognition processor with on-chip learning is fabricated in TSMC 65nm CMOS (Fig. 7, Table I). The object recognition processor runs at a maximum frequency of 635MHz at 1.0V and room temperature to achieve a high throughput of 10.16G pixel/s, dissipating 268mW (Fig. 5). The results demonstrate 8.2× higher throughput and 6.7× better energy efficiency than the previous SAILnet IM [4]. An example of recognizing an object is shown in Fig. 3. Tested with the MNIST database of 28×28 handwritten digits [6], the chip is capable of recognizing 9.9M objects/s at an accuracy of 84%. Increasing the inference period from 2τ to 12τ improves the classification accuracy to 90%, but cuts the throughput by 6×. The classification accuracy of this single-layer IM and single-layer classifier is still lower than what is reported in state-of-the-art machine learning literature, but the scalable architecture allows multiple layers of IM and classifier to be integrated in future work to improve the results. The on-chip learning co-processor runs at a maximum frequency of 650MHz at 1.0V, dissipating 258mW. A rigorous training using 1M image patches can be completed within three minutes. After learning converges, the co-processor is powered off.

The neuromorphic IM is error tolerant, and integrating IM and classifier provides additional error tolerance as the soft classifier accommodates more errors in feature extraction. Error-free classification can be achieved at a 450mV datapath supply and 425mV memory supply to improve the energy efficiency to 5.7pJ/pixel at 40MHz (Fig. 6). Compared to state-of-the-art neuromorphic ASICs, this design demonstrates enhanced capabilities and energy efficiency (Table II).

Acknowledgements

The work was supported in part by DARPA under cooperative agreement HR0011-13-2-0015. We thank W. Lu, M. Flynn, and G. Kenyon for suggestions.

References

[1] Rozell *et al*., *Neural computation*, vol. 20, no. 10, pp. 2526-2563, 2008.
[2] Lowe, *ICCV*, 1999.
[3] Shapero *et al*., *Neural Networks*, vol. 45, pp. 134-143, 2013.
[4] Kim *et al*., *VLSI Symp*., 2014.
[5] Mairal *et al*., *IEEE TPAMI*, vol, 34, no. 4, pp. 791-804, 2012.
[6] MNIST database. [Online]. http://yann.lecun.com/exdb/mnist.
[7] Seo *et al*., *CICC*, 2011.
[8] Merolla *et al*., *CICC*, 2011.

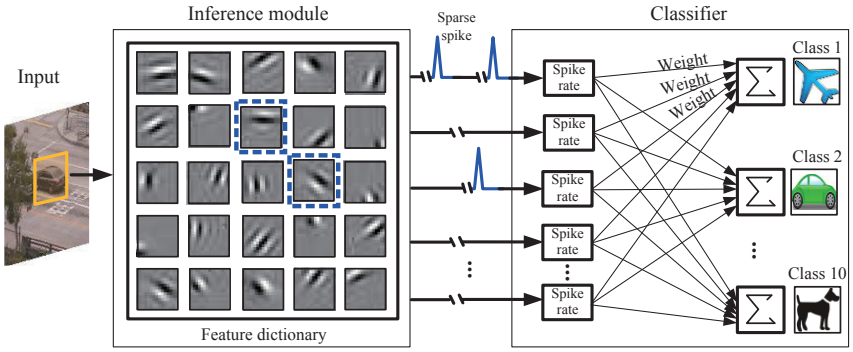

Fig. 1. Sparse neuromorphic object recognition system composed of the spiking LCA inference module (IM) front-end and the task-driven classifier back-end. A sparse set of features are extracted to represent the input image. The weighted spiking rate is summed to vote the most likely object class.

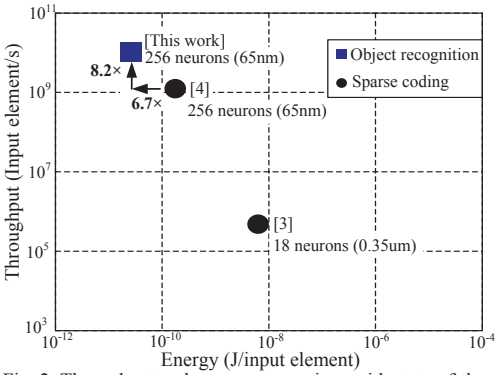

Fig. 2. Throughput and energy comparison with state-of-the-art neuromorphic ASICs for sparse coding.

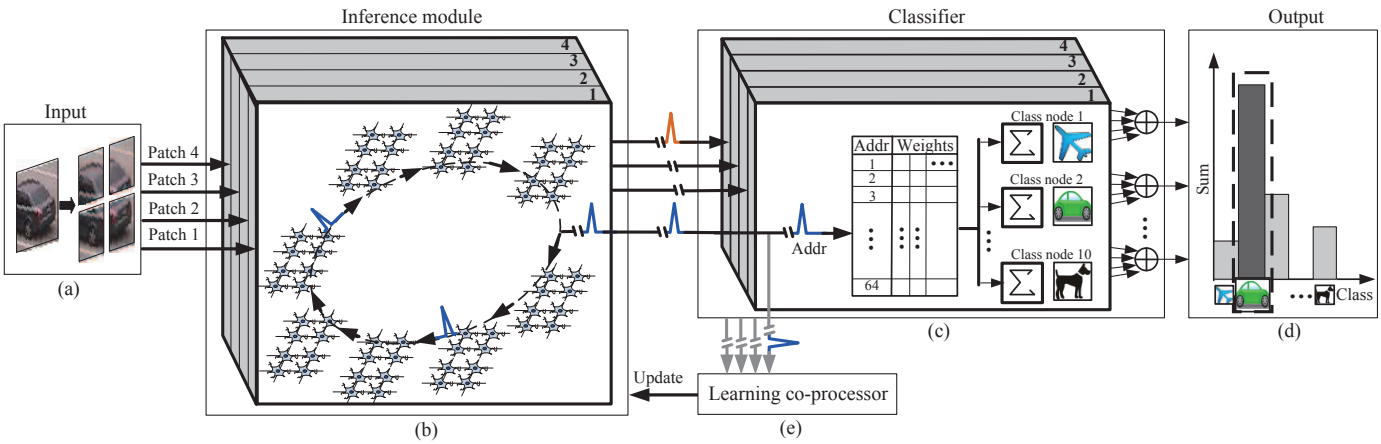

Fig. 3. Object recognition processor with on-chip learning co-processor. (a) Four image patches to extract features in parallel. (b) Inference module (IM) implemented in four 64-neuron spiking neural networks. (c) Spike event-driven classifier (d) Soft output of ten class nodes (e) On-chip learning co-processor.

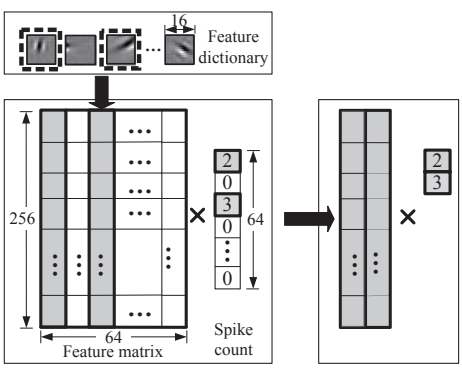

Fig. 4. (a) Feature matrix and a 64-entry spike count vector multiplication to support learning. (b) Simplified vector-matrix product by taking advantage of sparsity.

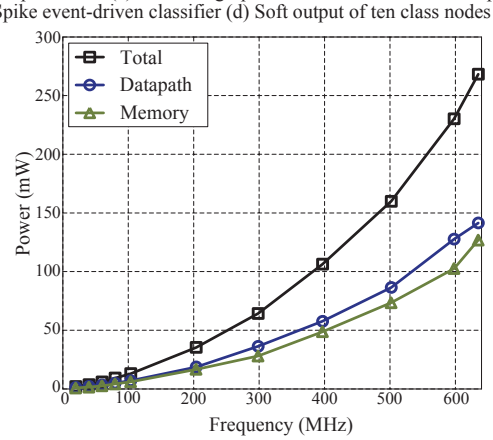

Fig. 5. Measured power consumption of the object recognition processor at the minimum datapath and memory supply voltages for each frequency.

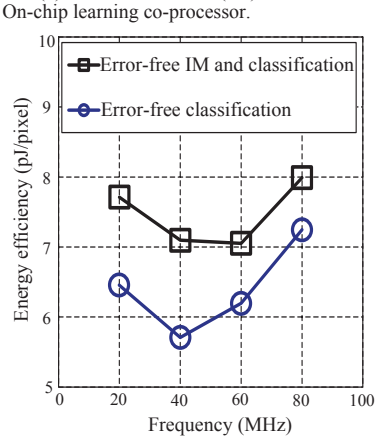

Fig. 6. Measured energy efficiency of the object recognition processor by exploiting error tolerance.

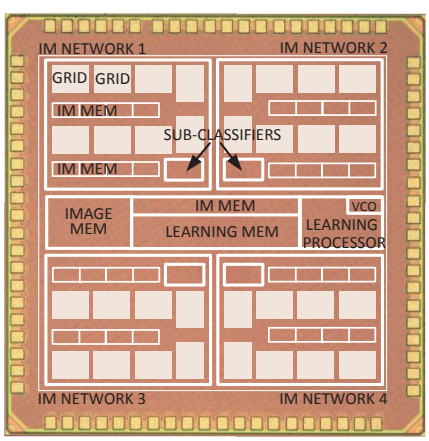

Fig. 7. Chip microphotograph

TABLE I: CHIP SUMMARY

Core Area	1.35mm × 1.35mm (Datapath : 0.97mm^2, Memory : 0.48mm^2, Learning : 0.21mm^2, Periphery: 0.16mm^2)	
Chip Area	1.73 × 1.73mm (2.99mm^2)	
Frequency (MHz)	40	635
Datapath (V)	0.45	1.00
Memory (V)	0.425	1.00
Throughput (Mpixel/s)	640	10160
Power (mW)	3.65	268.2
Energy Efficiency (pJ/pixel)	5.70	26.40

TABLE II: COMPARISION WITH PRIOR WORKS

Reference	Seo [7]	Merolla [8]	Shapero [3]	Kim [4]	This work
# Neurons	256	256	18	256	256
# Synapses	64K	256K	0.53K	128K	83K
Bitwith of a Synapse	4 bits	1 bit	7 bit	8 and 13 bits	4, 5 and 14 bits
Mem size	256Kbits	256Kbits	3.7 Kbits	1.31Mbits	301Kbits
Architecture	Crossbar	Crossbar	Crossbar	2-layer grid and ring	2-layer grid and ring
Algorithm	STDP	RBM	Spiking LCA	SAILnet	Spiking LCA with classification
Learning	On chip	Off chip	Off chip	On chip	On chip
Technology	45nm	45nm	0.35um	65nm	65nm
Core area	4.2mm^2	4.2mm^2	-	3.1mm^2	1.8mm^2
Energy metric	-	45pJ/spike	6.3nJ/input	48pJ/pixel	5.7pJ/pixel