A 127mW 1.63TOPS Sparse Spatio-Temporal Cognitive SoC for Action Classification and Motion Tracking in Videos

Ching-En Lee, Thomas Chen, Zhengya Zhang
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor

Abstract

A sparse spatio-temporal (ST) cognitive SoC is designed to extract ST features from videos for action classification and motion tracking. The SoC core is a sparse ST convolutional auto-encoder that implements recurrence using a 3-layer network. High sparsity is enforced in each layer of processing, reducing the complexity of ST convolution by two orders of magnitude and allowing all multiply-accumulates (MAC) to be replaced by select-adds (SA). The design is demonstrated in a 3.98mm^2 40nm CMOS SoC with an OpenRISC processor providing software-defined control and classification. ST kernel compression is applied to reduce memory by 43%. At 0.9V and 240MHz, the SoC achieves 1.63TOPS to meet the 60fps 1920×1080 HD video data rate, dissipating 127mW.

Introduction

Spatio-temporal receptive fields (STRFs) are understood as features or basis functions of videos [1]. STRFs can be extracted by unsupervised learning using an auto-encoder. Due to the high redundancy in video data, a compressed video encoding can be obtained using a sparse spatio-temporal (ST) auto-encoder [2]. This neuro-inspired approach provides not only efficient video coding but also cognitive processing capabilities such as action classification and motion tracking [3], [4] (Fig. 1).

We present a sparse ST convolutional auto-encoder SoC chip for video cognitive processing. The auto-encoder is realized in a network of neurons, each storing a 3D STRF (ST kernel) that is essentially a sequence of 2D features. Neurons compete to represent an input in 3 stages of iterative processing: 1) charge: each neuron performs ST convolution of its ST kernel with an input and charges up; 2) compete: neurons compete, and the potentials of non-active neurons are discharged; and 3) activate: the neurons generate spikes that are fed back to implement recurrence. The ST convolution accounts for the most expensive part of the processing, presenting challenges for a high-throughput implementation. In this work, we follow neuro-inspired principles to transform all steps of the processing to produce sparse spikes, and compute using sparse spikes to obtain over two orders of magnitude improvement in performance and efficiency over the baseline design.

Spatio-Temporal Cognitive SoC

The core of the SoC chip is a sparse ST convolutional auto-encoder that consists of 192 neurons, with each supporting a kernel up to 6×6×8 (6×6 frame, spanning 8 time steps) (Fig. 2). The auto-encoder is configurable with several settings: 64, 128 or 192 neurons, frame size from 1 to 36 and time steps from 1 to 8. Inputs are streamed in to the frame load queue, and ST kernels are reconstructed from their compressed storage prior to performing ST convolutions. The core is integrated with memory and an OpenRISC processor through a common control bus. The OpenRISC processor is programmed by an ISA together with a configuration and a classifier profile. The configuration profile controls the operation of the core during runtime; and the classifier profile configures the on-chip classification algorithm. The outputs of the core are sent to a communication hub in the OpenRISC processor.

Sparse Recurrent Network Architecture

Sparsity is often enforced by an ℓ_1 normalization term as part of the cost function in reference auto-encoder designs [1], [2] (Fig. 3(a)). To achieve an even higher sparsity, we reformulate the auto-encoder as a 3-layer recurrent network (Fig. 3(b)), and introduce ℓ_1 normalization in two layers using rectification (Fig. 4): 1) in Layer 1 (L_1), neurons compute ST convolutions to compute the recurrence and apply min/max rectification (i.e., hard thresholding to binary levels) to enforce a sparse spike rate of S_1, reducing the downstream workload by a factor up to $1/S_1$; 2) in Layer 2 (L_2), neurons compute ST convolutions to compute the potential update; and 3) in Layer 3 (L_3), neuron potentials are thresholded to generate sparse spikes at a target rate of S_3. The spikes are fed back to L_1, reducing L_1's workload by a factor up to $1/S_3$.

The three layers are fully parallelized using 192 neurons in each layer. Each L_1 and L_2 neuron performs a 6×6×8 ST convolution at a time, and each L_3 neuron updates its potential and performs thresholding. The number of iterations through the 3 layers is adjustable between 2 to 32 to meet processing requirements. To achieve a high classification accuracy while maintaining a low-power operation, the sparsity targets S_1 and S_3 are set to 3% and 1% respectively (Fig. 4), i.e., 97% and 99% of the L_1 and L_3 outputs are zero, enabling significant complexity and power reduction. In one iteration, the combined L_1 and L_2 workload is reduced to only 1 to 3% of the equivalent 3.54M OPs (an OP is defined as an equivalent 8b MAC) for a 6×6×64 (6x6 frame, 64 time steps) input video patch.

Spike-Based Inference and Sparsity-Enabled Compression

Spike inputs to L_1 and L_2 simplify L_1 and L_2 neuron implementation from multiply-accumulates (MAC) to select-adds (SA) triggered by sparse spikes (Fig. 5(a), (b)), thereby reducing neuron's power and area by 8.3× and 10.1× respectively. Spikes are detected by successively ANDing the spike train with its two's complement, which returns the one-hot encoding of the spike locations, i.e., the addresses of the ST kernel memory to read. In the absence of spikes, an entire layer will be skipped, enabling an average 3.5× power reduction and 6.3× latency reduction. Dynamic clock gating is enabled by the OpenRISC processor based on the configuration profile to cut the dynamic power by 4.2× when switching from 192 to 64 neurons to adapt to problem requirements.

The spike outputs of an L_1 neuron are aggregated over 8 time steps to reduce dimensionality (Fig. 5(d)), which is equivalent to a pooling operation in the time domain to compress data and reduce the latency of downstream processing. An L_3 neuron's outputs are encoded using the compressed column storage format (Fig. 5(c)). Due to sparsity, the compression results in 64 to 84% reduction in intermediate data storage.

ST kernels are quantized to 8 bits (Fig. 6(a)), and their storage requires 108KB, occupying 2.5mm^2 area in 40nm CMOS. We observe that the pixel value difference for 95% of the time-adjacent ST kernels vary within ±4 LSB (Fig. 6(b)). Therefore, we apply non-uniform delta coding (Fig. 6(c)) to compress ST kernels to 4 bits to reducing memory usage by 47.25KB and chip area by 43%. Prior to an ST convolution, ST kernels are reconstructed by a tree generator (Fig. 6(a)).

Chip Measurement and Classification Results

A 3.98mm^2 sparse ST cognitive SoC chip (Fig. 7) is implemented in 40nm CMOS. The chip achieves an effective 1.63TOPS with 0.9V supply at 240MHz. The performance meets the 60fps 1920×1080 HD video data rate, while dissipating 127mW (Fig. 8). The 6-class KTH human action dataset [5] is used for action classification testing (600 samples with train/test split ratio of 5:1). With the auto-encoder extracting the activation response of ST kernels, a softmax classifier implemented on the OpenRISC processor achieves a 76.7% classification accuracy. Using the same auto-encoder outputs, an off-chip SVM achieves an 82.8% accuracy (Table I). Motion tracking is also prototyped using a simple bounding box regression method based on the auto-encoder outputs. Compared to state-of-the-art vision processors [6], [7], this design offers enhanced capabilities of action classification and motion tracking using a recurrent network. The design exploits sparse spikes to effectively reduce workload, demonstrating competitive performance and efficiency (Table II). The sparse spatio-temporal SoC is suitable for a range of cognitive processing tasks.

Acknowledgements

This work is supported by DARPA UPSIDE, SONIC, Intel, and NSF GRFP.

References

[1] B. Olshausen, *IP Intl. Conf.*, 2003.
[2] M. Baccouche, et al., *BMVC*, 2012.
[3] S. Savarese, et al., *WMVC,* 2008.
[4] K. Zhang, et al., *ECCV,* 2012.
[5] C. Schüldt, et al., *ICPR*, 2004
[6] A. Suleiman, et.al., *VLSIC*, 2016.
[7] P. Knag, et. al., *VLSIC*, 2016.

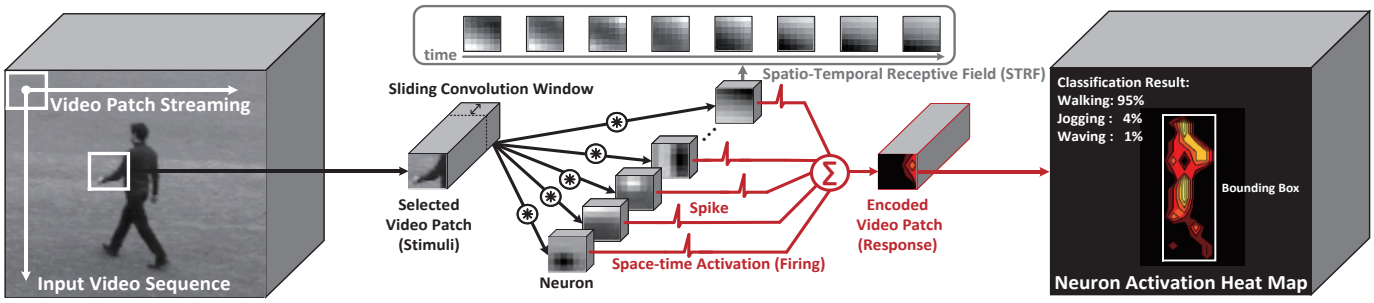


Fig. 1. Cognitive processing of video input using spatio-temporal (ST) convolutional auto-encoder for human action classification and motion tracking.

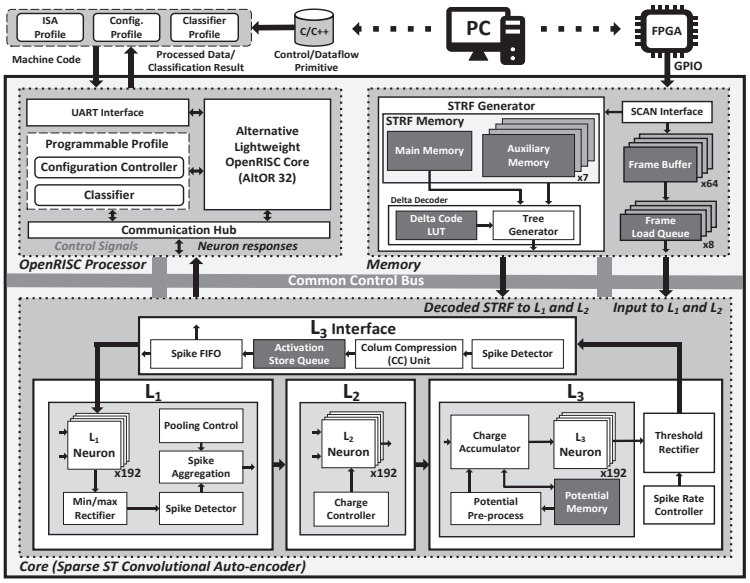


Fig. 2. Sparse spatio-temporal (ST) cognitive SoC system architecture, including an OpenRISC processor, memory, and a sparse ST convolutional auto-encoder (core).

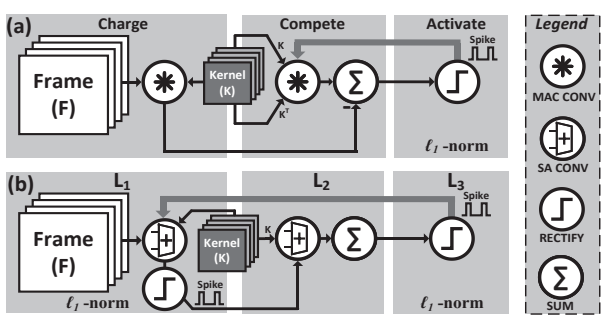


Fig. 3. (a) Baseline 3-stage charge-compete-activate inference architecture, (b) Modified 3-layer recurrent inference architecture.

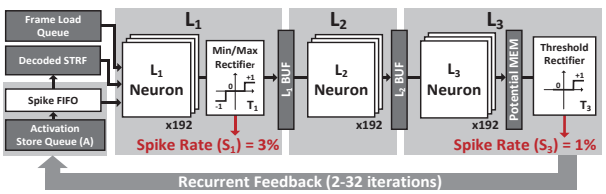


Fig. 4. Spatio-temporal (ST) auto-encoder (core) implemented in a 3-layer recurrent network, consisting of configurable sparsity of outputs at different layers, and configurable feedback iterations.

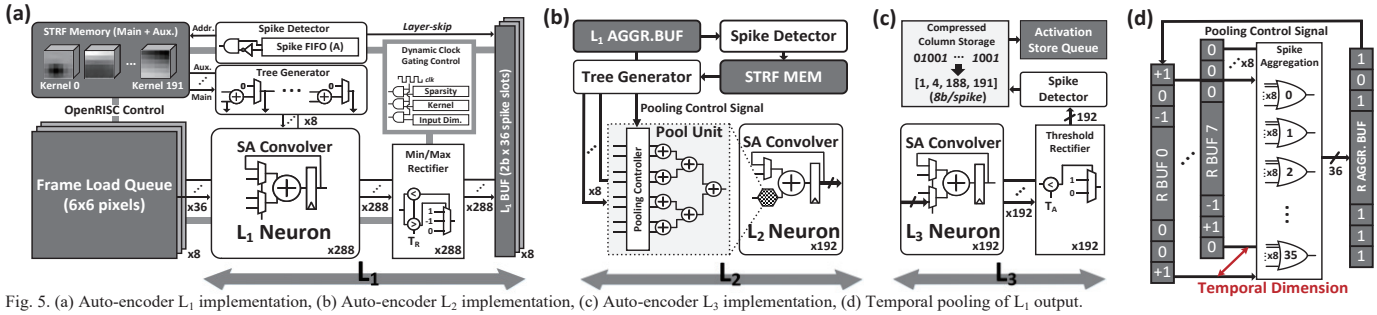
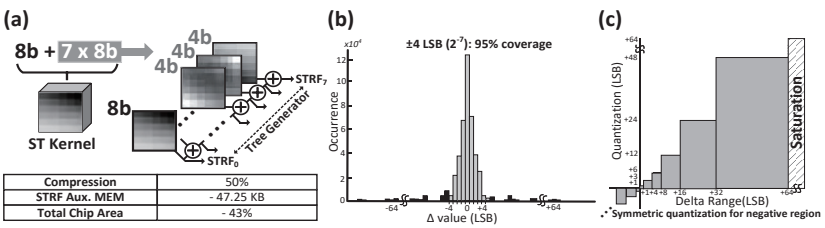


Fig. 5. (a) Auto-encoder L_1 implementation, (b) Auto-encoder L_2 implementation, (c) Auto-encoder L_3 implementation, (d) Temporal pooling of L_1 output.



Compression	50%
STRF Aux. MEM	- 47.25 KB
Total Chip Area	- 43%

Fig. 6. (a) Time-adjacent ST kernels are compressed to 4 bits, and reconstructed by a tree generator, (b) Histogram plot of all pixel value deltas for time-adjacent ST kernels, (c) Non-uniform delta coding quantization diagram.

TABLE I: COMPARISON OF CLASSIFICATION ACCURACY

Classes (individual classification accuracy)							
Algorithm	Box	Clap	Wave	Jog	Run	Walk	Total
On-Chip Softmax	70.0%	68.4%	85.0%	73.7%	94.4%	70.0%	76.7%
Off-Chip SVM	85.0%	78.9%	85.0%	73.7%	94.4%	80.0%	82.8%

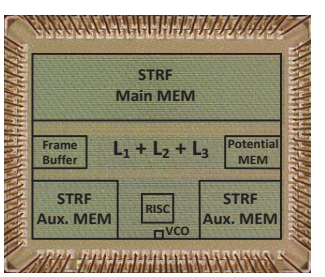


Fig. 7. Packaged chip microphotograph.

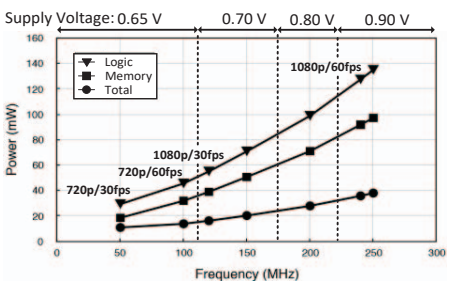


Fig. 8. Measured power and frequency at room temperature.

TABLE II: COMPARISON WITH PRIOR WORK

Reference	VLSIC'16 Suleiman [6]	VLSIC'16 Knag [7]	This Work
Application	Multi-Object Detection	Object Recognition	Action Classification Motion Tracking
Topology	Deformable Parts Model	Deep Neural Network	Sparse Convolutional Auto-Encoder
Technology	65 nm	40 nm	40 nm
Area	16.0 mm^2	1.4 mm^2	3.98 mm^2
Voltage	0.77 – 1.11 V	0.65 – 0.90 V	0.65 – 0.90 V
Frequency	62.5 – 125 MHz	120 – 240 MHz	50 – 250 MHz
Power	58.6 – 216.5 mW	40.9 – 140.9 mW	29.2 – 134.93 mW[(a)]
Frame Rate[(b)]	30 – 60 fps	N/A – 30 fps	30 – 60 fps
TOPS[(b)]	0.068 – 0.137	0.449 – 0.898[(c)]	0.815 – 1.630[(d)]
TOPS/W[(b)]	1.169 – 0.624	10.98 – 6.37[(c)]	14.818 – 12.835[(d)]

(a) Power is 127mW at 240MHz (60fps 1920x1080p HD video data rate), (b) Frame size is 1920x1080p HD video, (c) 1 OP is defined as an 8b multiply or a 16b add, (d) 1 OP is defined as an equivalent 8b multiply-accumulate (MAC).