Intelligent Vision Systems: Exploring the State-of-the-Art and Opportunities for the Future

Siddharth Advani*, Srinidhi Kestur*,‡, Vijaykrishnan Narayanan*
*The Pennsylvania State University, USA
‡Broadcom Corporation

***Abstract*— Vision and Video applications are becoming ubiquitous in mobile and embedded systems. The advent of wearable devices which require capabilities for real-time video analytics and prolonged battery lifetimes is further driving the need for innovative system designs with low-power, reliability and high performance. Further, the increasing resolution of image sensors in these mobile systems places an increasing demand on both the memory storage as well as the computational power. Such stringent requirements have given rise to accelerator-rich architectures in system-on-chips, where the primary computational burden is handled by dedicated hardware accelerators.**

In this paper we explore existing Vision accelerators and analyze their architecture, performance and scalability for different datasets and applications. The applications evaluated in this work are neuro-biologically inspired algorithms for object detection, object recognition and activity recognition which are complex, compute-intensive and bandwidth-intensive. This paper further analyzes the reliability of such embedded vision systems in terms of robustness of performance and energy efficiency under different application scenarios. Specifically, this work discusses the opportunities to improve energy efficiency by minimizing DRAM refreshes and explores techniques to exploit algorithmic resilience to minimize power consumption while maintaining reliable system accuracy and performance.

I. Introduction

Many chip-makers are now earmarking a significant amount of research effort for vision-based processors. Texas Instruments offers a heterogenous multi-core DSP for real-time vision applications using their Keystone architecture. Recently Freescale Semiconductor unveiled a vision system-on-chip - S32V - for accident-free-cars. Camera-friendly wearable devices like Google Glass are demanding better power efficiencies, improved performance and more powerful capabilities from the underlying technologies.

In the context of real-time vision applications, single-class object detection is a highly computationally intensive task. To robustly detect an object in an image that may appear at arbitrary position and scale involves (1) extracting optimized features that aptly describe the object and (2) searching the image in a sliding window fashion for the presence of particular configurations of the features that are indicative of the object's presence. This exhaustive search is compounded by objects that exhibit high appearance variability in shape, color and size. But, for visual-assist systems, the ability to perform such a task is imperative. For example, in a visual driving assist system, an approaching vehicle or a passing pedestrian needs to be detected with minimal latency, minimum false positives, and maximum accuracy. On the other hand, a wearable visual prosthesis device needs to augment the visual cognition of the user in diverse and vastly unconstrained environments for extended periods of time.

In this paper, we focus on xyz. To augment the next generation of wearables, we lay emphasis on abc. The main contributions of this paper are:

- We survey the state-of-the-art.
- We exploit reliability.

The rest of this paper is organized as follows: In Section II, we provide an overview of vision-based architectures and the corresponding state-of-the-art. Section III describes a robust object recogntion pipeline. Finally, we conclude with Section IV.

II. Related Work

Due to the capacity of human vision systems for highly complex processing at very low power, many brain-inspired algorithms and architectures have been proposed to emulate the human visual cortex. [1], [2], [3].

Even though Convolutional Neural Networks (CNNs) were explored in the early 1990s for vision applicationsi [4], they have resurfaced again after a long hiatus and become extremely popular in the past couple of years. This successful comeback can be attributed to two major phenomena: (1) the existence of large amount of data (needed to train the network well) with the evolution of the digital era, and (2) the development of custom hardware (required for acceleration) now being used for CNNs.

In the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) conducted in 2012, the winning team trained a CNN consisting of five convolutional and three fully-connected layers. Importantly, the depth of the CNN is critical to its recognition capabilities since the authors found that removing any convolutional layer resulted in inferior performance [5].

More recent and advanced CNN architectures have 10 to 20 layers of Rectified Linear Units, hundreds of millions of weights, and billions of connections between units. The reader is pointed to [6] for insights on deep architectures in general and [7] for CNN-based learning and their recent advances.

From a systems perspective, [8] mapped an earlier Convolutional Network based face-detection task onto custom hardware,

Most works in this domain have focused mainly on enhancing the performance and energy efficiency of the computational fabrics and do not address the inefficiencies of the main memory system. The memory system contributes between 10-30% of the overall power of embedded video systems and mobile phones [9]. The increasing memory size in new generations of embedded systems and the use of stacked 3D architectures that increase on-chip temperatures have drawn increasing attention on reducing the memory refresh energy. Consequently, there have been sustained efforts to introduce new power-efficient techniques such as Low Power Auto Self Refresh, Temperature Controlled Refresh, Refresh Pausing, Fine Granularity Refresh and Data Bus Inversion in new memory standards such as DDR4 [10]. Tuning DRAM refresh based on the data characteristics has been proposed as early as 1998 [11]. Recently, a software approach, termed as *Flikker* was proposed that relies on the user to annotate critical and non-critical parts [12]. It also allows refresh rates to be different for critical and non-critical sections of the memory and conserves the refresh energy.

III. Reliability

Reliability is being explored at different layers of abstraction; from devices [13], [14], [15] to memory [16] to algorithms.

In this section, we discuss the key design points of the pipeline. Figure 1 illustrates the end-to-end system beginning with the PCIe host data interface.

IV. Conclusion

In this work, we showcase alpha,betta,gamma

Future work entails uvw.

Acknowledgements

This work is supported in part by NSF Expeditions: Visual Cortex on Silicon CCF 1317560. The work is also supported through infrastructure provided by NSF Award 1205618.

References

[1] A. Nere, A. Hashmi, and M. Lipasti, "Profiling Heterogeneous Multi-GPU Systems to Accelerate Cortically Inspired Learning Algorithms," in *IPDPS*, 2011.

[2] T. Chen, J. Wang, Y. Chen, and O. Temam, "DianNao : A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning," in *ASPLOS*, 2014.

[3] S. Kestur *et al.*, "Emulating Mammalian Vision on Reconfigurable Hardware," in *IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, April 2012.

[4] S. Lawrence, C. Giles, and A. C. Tsoi, "Convolutional Neural Networks for Face Recognition," in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, Jun 1996, pp. 217–222.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[6] Y. Bengio, "Learning Deep Architectures for AI." Now Publishers, 2009.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," in *Nature*, May 2015, pp. 436–444.

[8] C. Farabet, C. Poulet, J. Han, and Y. LeCun, "CNP: An FPGA-based processor for Convolutional Networks," in *International Conference on Field Programmable Logic and Applications (FPL)*, Aug 2009.

[9] A. Carroll and G. Heiser, "An Analysis of Power Consumption in a Smartphone," in *Usenix Annual Technical Conference*, 2010.

[10] "JEDEC DDR3 and DDR4 SDRAM Standard," 2012.

[11] T. Ohsawa, K. Kai, and K. Murakami, "Optimizing the dram refresh count for merged DRAM/logic LSIs," in *ISLPED*, 1998.

[12] S. Liu, Pattabiraman K., T. Moscibroda, and B. Zorn, "Flikker: Saving DRAM Refresh-power through Critical Data Partitioning," in *ASLPOS*, 2011.

[13] S. Datta, H. Liu, and V. Narayanan, "Tunnel FET technology: A reliability perspective," *Microelectronics Reliability*, vol. 54, no. 5, pp. 861 – 874, 2014.

[14] N. Agrawal, H. Liu, R. Arghavani, V. Narayanan, and S. Datta, "Impact of Variation in Nanoscale Silicon and Non-Silicon FinFETs and Tunnel FETs on Device and SRAM Performance," *Electron Devices, IEEE Transactions on*, vol. 62, no. 6, pp. 1691–1697, June 2015.

[15] R. Pandey *et al.*, "Tunnel Junction Abruptness, Source Random Dopant Fluctuation and PBTI Induced Variability Analysis of $GaAs_{0.4}Sb_{0.6}/In_{0.65}Ga_{0.35}As$ Heterojunction Tunnel FETs," *IEEE International Electron Devices Meeting*, (Accepted) 2015.

[16] L. Chen and Z. Zhang, "MemGuard: A low cost and energy efficient design to support and enhance memory system reliability," in *Computer Architecture (ISCA), 2014 ACM/IEEE 41st International Symposium on*, June 2014, pp. 49–60.

Fig. 1. System Architecture mapped to an FPGA