Visual Cortex on a Chip

Large-scale, real-time functional models of mammalian visual cortex on a GPGPU

Steven P. Brumby, Amy E. Galbraith, Michael Ham, Garrett Kenyon, and John S. George
Los Alamos National Laboratory, Los Alamos, New Mexico, USA

Goal
Los Alamos National Laboratory’s Petascale Synthetic Visual Cognition project is exploring full-scale, real-time functional models of human visual cortex using the Roadrunner petaflop (1000 teraflop) supercomputer and future GPGPU-based exaflop (1000 petaflop) computers. The project’s goal is to understand how human vision achieves its accuracy, robustness and speed. Commercial-off-the-shelf hardware for this modeling is rapidly improving, e.g., a teraflop GPGPU card for a workstation now costs ~$500 and is ~size of mouse cortex. We now present initial results demonstrating whole image classification using standard computer vision image datasets, and object extraction from UAV video using a model of primary visual cortex running on a GPGPU (240-core NVIDIA GeForce GTX285). As this technology continues to improve, cortical modeling on GPGPU devices has the potential to revolutionize computer vision.

Functional Models of Visual Cortex. Processing in the human visual system starts in the retina of the eye, continues in the brain until the retinotopic nucleus (LGN), and then reaches the cortex at primary visual cortex (V1). The visual processing starts in the retina and continues in the LGN and then reaches cortex through layers of cortical neurons operating in functional groups. Nuclei and thalamus model V1 consists of several layers: S-cells and complex “C” cells [1]. Fukushima and Poggio et al. [3-6] have proposed hierarchical models of the extrastriate pathway (PIT pathway) comprised of visual cortical regions V1, V2, V4, supporting a model of whole object detection in interhemispheric cortex (IT).

Model
LANL’s Petascale Artificial Neural Network (PANN) [8] is a high performance C++, C, and Python implementation of a feedforward-type hierarchical model of human visual cortex regions V1 (primary visual cortex), V2, V4, and interhemispheric cortex (IT). The ventral pathway of visual processing ("face pathway") within PANN exploits conventional clusters of multi-core CPUs or hybrid machines such as IBM Cell-based clusters (Roadrunner architecture [9]), or GPGPU-accelerated clusters that are currently in development. Pinto, Cox & DiCarlo [7] have recently shown high-throughput image and video processing with V1-type models using a multi-GPU architecture.

The key scientific question is how does visual cortex organize itself in response to large amounts of visual stimuli? PANN is designed to process large amounts of still and video imagery to match the visual input to the brain. PANN processes this imagery using unsupervised learning algorithms to build a hierarchical representation of natural scenes, combined with a relatively small amount of supervised learning required to train a back-end classifier (e.g., linear kernel SVM or neural network). The network is trained to be sparse, drawing from a reductionist dictionary of scene elements (DiCarlo & Park, [12]), which is driving development of a rich set of new ideas about pattern recognition. The speed achievable by running on clusters of GPGPU-accelerated compute nodes will enable testing of the properties of some of these learning rules at the full scale of human visual experience (~1015 images/year).

Results
We ported our PANN C++ code to NVIDIA GPGPU using CUDA to develop C++ host code and C device code. The porting process was straightforward and fast to complete. The PANN code is optimized for ease of algorithm exploration, e.g., we use global device memory to store neuron columns for a mechanism of pattern recognition unaffected by shift in position, and we also exploit the CUDA platform to store neuron columns for a mechanism of pattern recognition unaffected by shift in position. We compared performance against a multi-core CPU Linux cluster [10] and found significant speed-ups across the model.

We compare speed of a single-core PANN model using 32-bit arithmetic running on 2 x G1 Tesla single-GPU servers to multi-GPU servers. We compared performance against a multi-core CPU Linux cluster [10] and found significant speed-ups across the model. We found that the C-cell module, which is the most computationally intensive part of the model, obtained a 3x speedup. The C-cell module shared host CUDA computing. We also found that a small fraction of the code running on NVIDIA GPUs can be parallelized with host computing. Hence, we conclude that GPUs with CUDA enable significant visual cortex model speed-up without significant constraints on algorithm exploration.

Necessary computing hardware is now available. LANL’s Roadrunner supercomputer recently reached a petaflop [11] marking the arrival of computing platforms large enough and fast enough for full-scale, real-time functional modeling of human cortex [7]. However, small mammals are capable of excellent visual acuity and object recognition with brains orders of magnitude smaller than human (11G cortical neurons in human vs 3G in cat or 0.004G in mouse). GPGPU technology could enable widespread use of such small-scale, real-time models of mammalian visual cortex for a wide range of computer vision tasks.

References