Chapter 4.19
Stream Processing of a Neural Classifier I

M. Martínez-Zarzuela
University of Valladolid, Spain

F. J. Díaz Pernas
University of Valladolid, Spain

D. González Ortega
University of Valladolid, Spain

J. F. Díez Higuera
University of Valladolid, Spain

M. Antón Rodríguez
University of Valladolid, Spain

INTRODUCTION

An *Artificial Neural Network* (ANN) is a computational structure inspired by the study of biological neural processing. Although neurons are considered as very simple computation units, inside the nervous system, an incredible amount of widely inter-connected neurons can process huge amounts of data working in a parallel fashion. There are many different types of ANNs, from relatively simple to very complex, just as there are many theories on how biological neural processing works. However, execution of ANNs is always a heavy computational task. Important kinds of ANNs are those devoted to pattern recognition such as *Multi-Layer Perceptron* (MLP), *Self-Organizing Maps* (SOM) or *Adaptive Resonance Theory* (ART) classifiers (Haykin, 2007).

Traditional implementations of ANNs used by most of scientists have been developed in high level programming languages, so that they could be executed on common *Personal Computers* (PCs). The main drawback of these implementations is that though neural networks are intrinsically parallel systems, simulations are executed on a *Central Processing Unit* (CPU), a processor designed for the execution of sequential programs on a *Single Instruction Single Data* (SISD) basis. As a result, these heavy programs can take hours or even

DOI: 10.4018/978-1-60960-195-9.ch419

days to process large input data. For applications that require real-time processing, it is possible to develop small ad-hoc neural networks on specific hardware like *Field Programmable Gate Arrays* (FPGAs). However, FPGA-based realization of ANNs is somewhat expensive and involves extra design overheads (Zhu & Sutton, 2003).

Using dedicated hardware to do machine learning was typically expensive; results could not be shared with other researchers and hardware became obsolete within a few years. This situation has changed recently with the popularization of *Graphics Processing Units* (GPUs) as low-cost and high-level programmable hardware platforms. GPUs are being increasingly used for speeding up computations in many research fields following a *Stream Processing Model* (Owens, Luebke, Govindaraju, Harris, Krüger, Lefohn & Purcell, 2007).

This article presents a GPU-based parallel implementation of a Fuzzy ART ANN, which can be used both for training and testing processes. Fuzzy ART is an unsupervised neural classifier capable of incremental learning, widely used in a universe of applications as medical sciences, economics and finance, engineering and computer science. CPU-based implementations of Fuzzy ART lack efficiency and cannot be used for testing purposes in real-time applications. The GPU implementation of Fuzzy ART presented in this article speeds up computations more than 30 times with respect to a CPU-based C/C++ development when executed on an NVIDIA 7800 GT GPU.

BACKGROUND

Biological neural networks are able to learn and adapt its structure based on the external or internal information that flows through the network. Most types of ANNs present the problem of *catastrophic forgetting*. Once the network has been trained, if we want it to learn from new inputs, it is necessary to repeat the whole training process from the beginning. Otherwise, the ANN would forget previously acquired knowledge. S. Grossberg developed the *Adaptive Resonance Theory* (ART) to address this problem (Grossberg, 1987). Fuzzy ART is an extension of the original ART 1 system that incorporates computations from *fuzzy set theory* into the ART network, and thus making it possible to learn and recognize both analog and binary input patterns (Carpenter, Grossberg & Rosen, 1991).

GPUs are being considered in many fields of computation and some researchers have made efforts for integrating different kinds of ANNs on the GPU. Most research has been done for implementing *Multi-Layer Perceptron* (MLP) taking advantage of the GPU performance in matrix-matrix products (Rolfes, 2004) (Oh & Jung 2004) (Steinkraus, Simard & Buck 2005). Other researchers have used the GPU for *Self-Organizing Maps* (SOM) with great results (Luo, Liu & Wu, 2005) (Campbell, Berglund & Streit, 2005). Bernhard et al. achieved a speed increase of between 5 and 20 times simulating large networks of *Spiking Neurons* on the GPU (Bernhard & Keriven, 2006). Finally, Martínez-Zarzuela et al. developed a generic *Fuzzy ART* ANN on the GPU achieving a speed up higher than 30 over a CPU (Martínez-Zarzuela, Díaz, Díez & Antón, 2007).

Commodity graphics cards provide a tremendous computational horsepower. NVIDIA's GeForce 7800 GTX GPU is able to sustain 165 GFLOPS against the 25.6 GFLOPS theoretical peak for the SSE units of a dual-core 3.7 GHz Intel Pentium Extreme (Owens, Luebke, Govindaraju, Harris, Krüger, Lefohn & Purcell, 2007). Newest generation of graphics cards, like NVIDIA Geforce 8800 Ultra, or AMD (ATI) Radeon HD 2900 XT, can give a peak performance higher than 500 Gflops and 100 GB/s peak memory bandwidth. Graphics cards manufacturers have recently discovered the field of high performance computing as to be a target market for their products and are providing specific hardware and software to couple with enterprises and researchers heavy computational requirements.