BLOCK-BASED NEURAL NETWORK MAPPING ON GRAPHICS PROCESSOR UNIT

ONG CHIN TONG

UNIVERSITI TEKNOLOGI MALAYSIA

BLOCK-BASED NEURAL NETWORK MAPPING ON GRAPHICS PROCESSOR UNIT

ONG CHIN TONG

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Electrical-Computer and Microelectronic System)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

JUNE 2015

Dedicated, in thankful appreciation for support, encouragement and understanding to my beloved mother, father, brother and supervisor....

ACKNOWLEDGEMENT

First and foremost, thank God for giving me the strength to complete this project. I wish to express my greatest appreciation to my supervisor, Assoc Prof Dr. Muhammad Nadzir bin Marsono for his generous guidance, advice and motivation throughout this study. His critical suggestion and comments help me to develop the study project into understandable and workable effort.

Besides that, I would also like to thank Dr Vishnu P. Nambiar from Intel for providing me the suggestion of this topic. Thank you for willing to allocate his time to guide me throughout my works.

Finally, acknowledgements are also for my family, especially my parents. Thank you for their support, inspiration, and encouragement during the study and completion of this thesis.

Ong Chin Tong

ABSTRACT

Block-based neural network (BbNN) was introduced to improve the training speed of artificial neural network. Various works had been carried out by previous researchers to improve training speed of BbNN system. Multithread BbNN training on field-programmable gate array (FPGA) limits training speed due to low performance of Nios II software used for communication between central processing unit (CPU) and FPGA. This project aims to improve training speed of multithread BbNN block by mapping BbNN model into Compute Unified Device Architecture (CUDA) core. In this project, each BbNN block is mapped into a CUDA core with each core running on a single thread. The functional verification of BbNN core is carried out based on the BbNN output accuracy value. Near 100 percent accuracy value obtained is used to verify the CUDA mapped BbNN. The performance trade-off analysis had been carried out by comparing the accuracy value obtained from BbNN evolution on GPU versus CPU implementations. From the results obtained, it is found out that the performance of CUDA-mapped BbNN can only be as fast as CPU-mapped implementation. Although CUDA-mapped BbNN implementation run multiple BbNN blocks training in parallel, large data transfer between CPU and GPU dominates the performance gain in training multiple BbNN blocks in parallel. Besides that, a significant gain in training speed can only be seen if the order of complexity for GPU execution is at a higher order compared to the order of CPU-GPU data transfer. The result obtained in this project provides recommendation for future research works on how to further improve the training speed of CUDA-base BbNN implementation.

ABSTRAK

Block rangkaian neural (BbNN) telah diperkenalkan untuk menyingkatkan masa pemprosesan rangkaian neural. Pelbagai kerja telah dijalankan oleh penyelidik sebelum ini untuk menyingkatkan masa pemprosesan BbNN. Prestasi multithread BbNN menggunakan field-programmable gate array (FPGA) akan dihadkan oleh prestasi perlahan daripada perisian Nios II yang digunakan untuk berkomunikasi antara central processing unit (CPU) dan FPGA. Projek ini bertujuan untuk menerokai kaedah bagi menyingkatkan masa pemprosesan dengan memetakan BbNN mengunakan teras Compute Unified Device Architecture (CUDA). Dalam projek ini, setiap blok BbNN dipetakan ke dalam teras CUDA dengan setiap teras berjalan dengan satu thread. Dengan mendapat ketepatan yang hampir kepada 100 peratus, BbNN yang dipetakan ke dalam CUDA telah disahkan betul. Perbandingan antara prestasi GPU dan CPU kemudian dijalankan dengan mendapatkan perbezaan ketepatan dan masa pemprosesan BbNN. Daripada keputusan projek ini, didapati kelajuan pemprosesan BbNN yang dipetakan ke dalam teras CUDA hanya seiras dengan masa pemprosesan BbNN CPU. Walaupun BbNN yang dipetakan ke dalam teras CUDA diprocess secara selari, prestasi masa pemprosesan CUDA telah didominasi oleh jumlah besar data yang perlu disampaikan antara CPU dan GPU. Di samping itu, peningkatan prestasi pemprosesan CUDA hanya dapat diperlihat sekiranya kerumitan pembilangan berada dalam order yang lebih tinggi daripada kerumitan data yang perlu disampaikan. Keputusan yang diperolehi daripada projek ini dapat menyediakan cadangan untuk kajian masa hadapan mengenai cara untuk meningkatkan lagi prestasi BbNN yang dipetakan ke dalam teras CUDA.