Development of a Modular System for Hardware Acceleration of Machine Learning

This research proposal outlined in this document aims to provide a hardware solution for accelerating machine learning in a data centre environment. The initial aim would be to produce hardware on an FPGA that would allow for the acceleration of a number of different algorithms using System Verilog. This project will also aim to produce a modular system for generation of HDL code to implement the required functions.

Abstract

Contents

1 Motivation for Applying to MINDS CDT

My motivation for applying to MINDS CDT has been guided significantly by the aspects of my degree that I have enjoyed the most, which predominantly involve digital design, both using HDL and ASIC design. I have found that my Part 3 Project has been the single most enjoyable aspect of my degree, having involved a substantial amount of design using HDL. The MINDS programme is particularly interesting to me as it would allow me to apply my existing knowledge of HDL design for FPGAs to an area of which I have relatively little experience, machine learning. I would enjoy the challenge that this would provide, both in terms of testing my abilities in an area in which I am familiar and challenging me to develop new skills. As requested in the page detailing what is required in my application, I would like to declare my interest in becoming an ambassador.

2 Introduction

Machine learning is becoming increasingly widespread in the modern datacentre. In the datacentre, work has been directed towards increasing the speed with which algorithms can be applied to large datasets. Traditionally, algorithms have been accelerated using GPUs, however, FPGAs can offer improvements regarding both computational speed and power efficiency [1,2].

While an ASIC provides more options with regard to optimising the device's characteristics in various ways, this proposal will not consider the possibility of production of an ASIC, as the ability of an FPGA to be reconfigured is central to the aims of this project, in that it seeks to provide a modular system in which an FPGA can be quickly reconfigured to suit the needs of different machine learning algorithms.

The idea of a modular system based on the workload has already been proven in the case of the RISC-V processor architecture, this project would aim to produce a similar system for FPGA acceleration of machine learning, with hardware produced according to the needs of the user [3].

3 Research Aims

The aim of this research will be to create a software and hardware ecosystem that will have the capability to accelerate multiple machine learning algorithms. This will require a range of hardware capable of supporting the mathematical operations needed for machine learning.

Investigating what is required for multiple machine learning algorithms will lead to the development of a modular system which will examine what is required for a given workload and configure the FPGA with the appropriate modules. This allows for two advantages. By removing unnecessary hardware from the design, this allows for a configurable balance between power saving by having less hardware, and the inclusion of an increased number of structures which are used, allowing for increased parallelisation of a work load for improved performance. These are both important considerations in the data centre, which is the primary environment in which these devices would be used.

Since the accelerator will support multiple different algorithms, this project will aim to develop a software layer that abstracts control of the hardware away from the user. This would allow for developers to make use of hardware acceleration using the provided platform without requiring low level knowledge of the hardware. This software would be responsible for observing which functions are required for a given task and producing the required HDL code.

This research will aim to measure and quantify the performance increase and power consumption reduction against a system using GPU acceleration in order to validate that the proposed method is producing hardware that achieves the goal. Producing the design in a modular fashion also allows for the implementation of a design on a less costly FPGA as it allows for the usage of less resources compared to configuring the FPGA with all of the hardware for all possible algorithms.

4 Research Method

Initially, research would be directed towards identifying which stages of a given machine learning algorithm are the slowest processes, as addressing these areas will result in the greatest performance increase. This will provide information as to which functions are required to be implemented for each algorithm.

These would then be implemented as System Verilog modules that describe these various functions. This will allow for the beginning of development of the modular system. Due to the nature of some of the popular algorithms in machine learning, it is likely that some functions within the hardware are likely to be required in multiple different configurations. The functions that are likely to be replicated include those that perform tasks such as vector and matrix manipulation.

In order to maximise the capability of the accelerator to parallelise tasks, it will also be required to account for the number of logical elements required for each of the modules necessary for a given algorithm. This will allow for optimal usage of resources on the FPGA by including the maximum amount of hardware for any necessary functions. This will also require accounting for Amdahl's law. By reducing the time taken for the slowest process in the algorithm, other processes will become the slowest, hence the performance gain from improving this process will become limited. This will require finding a balance of hardware that produces the maximum performance increase across multiple processes. This will be determined by experimentation, by including different amounts of each of the functions and measuring the impact on performance.

In order to develop a software abstraction layer, it will be required for the software to determine which FPGA is being used, and which machine learning algorithms are being applied to a problem. From this information and knowledge of which functions are required to successfully implement an algorithm, HDL code should be assembled and used to configure the device. This is anticipated to be the most challenging area of the project, as it would require creating a tool that will semantically analyse code and be capable of producing HDL code that implements the key functionalities in hardware.

As previously mentioned, the produced system's performance will be compared to the traditional method of accelerating machine learning, which uses the compute capabilities of GPUs. The relative performance of the two methods will be measured by comparing their performance in example workloads that make use of typical machine learning techniques, for example, a computer vision task based around image recognition. The results would be analysed based on power consumption throughout the test and the time taken for completion of the given task. Ideally, the designed system would be tested against a traditional implementation of FPGA acceleration to establish whether this method of designing an acceleration system offers any improvement over the traditional design methods.

5 Significance of This Research

This research has the potential to be significant due to the widespread applications that this would have in the increasingly data driven world due to the growing need to process large amounts of data and find correlation within it. This project builds upon work previously done towards implementing FPGA acceleration for machine learning tasks by adding new possibilities to increase the amount of throughput of each device by decreasing the amount of unused resources on each FPGA and increasing the amount of hardware being used. The typically low power consumption of FPGAs also provides a better solution to the problem than using GPU accelerated methods. This also provides options regarding cost of the design due to the range of FPGAs available.

This solution also provides significantly greater flexibility than a solution based around ASICs, due to the reconfigurability of FPGAs. This solution would provide the ability to quickly make alterations to any algorithm implemented by changing the FPGA, while an ASIC, while potentially offering greater performance only enables one possible configuration of hardware. In contrast to this method, this has the potential to become obsolete more quickly, as the ASIC is unable to be changed to adapt to the

changing requirements caused by the development of new machine learning techniques and algorithms. Designing a system based upon an ASIC also requires any necessary hardware be included prior to fabrication. As a result of this, any hardware required for any of the supported algorithms must always be present on the IC. Designing the accelerator in a modular way would allow for a maximum of on chip resources to be dedicated to solving the problem at hand, without removing the ability to use the same hardware to solve other problems.

6 References

C. Wang, L. Gong, Q. Yu, X. Li, Y. Xie and X. Zhou, "DLAU: A Scalable Deep Learning Accelerator Unit on FPGA," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 36, no. 3, pp. 513-517, March 2017.
doi: 10.1109/TCAD.2016.258768310.1109/TCAD.2016.2587683

K. Freund, “An Update On Data Center FPGAs”, Forbes, July 20 2018, [Online] Available: https://www.forbes.com/sites/moorinsights/2018/07/20/an-update-on-data-center-fpgas/#346b170446c2, Accessed on April 21 2019

J. McGregor, “The Difference Between ARM, MIPS, x86, RISC-V And Others In Choosing A Processor Architecture”, Forbes, April 5 2018, [Online] Available: https://www.forbes.com/sites/tiriasresearch/2018/04/05/what-you-need-to-know-about-processor-architectures/#51831d774f57, Accessed on April 21 2019