DATX02-22-15

Machine Learning Accelerator Co-Design

Description

Machine Learning (ML) is an emerging application domain, which is very demanding on the hardware both in terms of computation and memory. A way to improve the performance for these applications is through the use of dedicated hardware (HW) accelerators (e.g. Google TPU).

When designing a HW accelerator, it is possible design computational units that satisfy the requests of the software. On the other hand, it is possible for the software (SW) to use algorithms and techniques that are aware of the underlying hardware. While these are two common approaches that can achieve some performance benefit, this is still not the best that can be achieved! To exploit the most benefit it is necessary to design together both the SW and the HW in a process known as co-design. A well known approach is to take both software and hardware blocks and try to experiment different combinations of these until the best co-design combination is found. The challenge in co-design is that the design space (possible options) for these co-designed accelerators is quite large as we can change at the same time both the hardware and the software.

The work in this project focuses on exploring aspects of the co-design for ML accelerators.

Goals

With this project we would like to learn:

- Which are some of the techniques used for co-design?
- Which software blocks are appropriate for the accelerator co-design?
- Which hardware blocks are appropriate for the accelerator co-design?
- What are efficient design space selection techniques to help in the accelerator co-design?
- How much benefit can this technique achieve in comparison with more traditional accelerator design?

Pre-requisites

For the success in this project the student needs to have basic knowledges of computer architecture. Basic knowledge of Machine-Learning is an added benefit.

Methodology

- Selection of a ML application – selection of ML model and platform (training of a model is not required, pre-trained models should be used when necessary)
- Survey of co-design techniques (read the literature)
- Problem formulation (finding an interesting research question)
- Depending on the interest of the student(s) focus on one or more of the following:
  - Development of interesting software blocks and how to decompose the original application into these blocks
  - Development of interesting hardware blocks and how to build an accelerator how of these blocks
  - Exploration of a selection mechanism for the different combinations of hardware and software blocks

- Evaluation of the proposed technique and comparison with tradition accelerators (there is the possibility to build a FPGA prototype for the HW blocks)

Selected Relevant References

- Cong Hao, Yao Chen, Xiaofan Zhang, Yuhong Li, Jinjun Xiong, Wen-mei Hwu, Deming Chen, " Effective Algorithm-Accelerator Co-design for AI Solutions on Edge Devices", arXiv:2010.07185 [cs.AR]
- Cong Hao et al. "FPGA/DNN co-design: An efficient design methodology for IoT intelligence on the edge". In Proceedings of the ACM/IEEE Design Automation Conference (DAC), 2019
- Xiaofan Zhang et al. SkyNet: a hardware-efficient method for object detection and tracking on embedded systems. In Proceedings of Machine Learning and Systems (MLSys), 2020

Target group

DV, D, E and IT

Contact:

Pedro Petersen Moura Trancoso
Computer Science and Engineering
ppedro@chalmers.se