Imperial College London

MEng Individual Project

Imperial College London

Department of Computing

Reconfigurable Acceleration of Transformer Neural Networks with Meta-Programming Strategies for Particle Collisions Experiments

Author:
Filip Wojcicki

Supervisor:
Prof. Wayne Luk

Second Marker:
Dr. TODO

January 25, 2022

Acknowledgements

Although the project is at an early stage, I would like to express my gratitude to Professor Wayne Luk and Zhiqiang Que for guiding me through the project and always being available to answer any of my questions.

Contents

Chapter 1

Introduction

1.1 Overview

Particle physics is one of the key branches of modern physics, with the Standard Model theory at its core. It tackles the underlying questions about the nature of the universe by describing the fundamental forces and elementary particles. In order to verify the correctness of the theories, countless experiments have to be designed and carefully executed, with the main driving force of myriads of engineers, physicists and researchers at Large Hadron Collider (LHC) operated by the European Organization for Nuclear Research (CERN).

LHC is the world's highest-energy particle collider that is capable of producing and detecting the heaviest types of particles that emerge from collisions such as a proton-proton collisions. The detection is a challenging process as some particles like quarks and gluons cannot exist on their own, and they nearly instantly combine which results in a collimated spray of composite particles (hadrons) that is typically referred to as a **jet** [1]. The initial particles created upon collision and their behaviors are of main interest of the physicists, which leads to **jet tagging** - the challenge of associating particle jets with their origin.

1.2 Motivation

There are many detector types used for the analysis the particle collisions, each based on a different physical methodology, which result in availability of both higher and lower level features. The former have been successfully used in the past using more physically motivated machine learning (ML) algorithms, e.g. using computer vision [2]. However, more recently, various deep learning approaches have proven to outperform their predecessors [3]. It has also been found that all the detected features carry the same underlying information, with convolutional neural networks (CNN) trained on higher-level data achieving nearly identical accuracy as dense neural networks (DNN) trained on the data from the other end of the spectrum [4].

The Pb/s throughput of information collected by the LHC detectors outclasses the real-time inference capabilities of the typical state-of-the-art solutions. The real-time decision-making is often required, hence this paper is motivated by the successful adoption of ***hls4ml*** codesign workflow in particle physics experiments [5]. It allows ML researchers and physicists to easily deploy their solutions trained using common ML frameworks on reconfigurable or application specific hardware, vastly improving the detection algorithms throughput. However, *hls4ml* lacks support for a number of neural network architectures that have been proven to outperform the previous state-of-the-art, including graph neural networks (GNN) [6, 7] and transformer neural networks [8].

1.3 Objectives and Challenges

The purpose of this project is to develop state-of-the-art neural network architectures for Field-Programmable Gate Arrays (FPGA) technology. While working towards this goal, there is an

emphasis on creating parametrizable and reusable designs as the next objective is to use metaprogramming strategies to integrate them into the *hls4ml* library with various optimizations that offer trade-offs between speed and hardware resources usage.

The two main challenges of the project involve:

- Developing deep and complex neural networks in hardware which requires working at a much lower abstraction level than a typical ML frameworks. It is also crucial to stay aware of the underlying hardware architecture to exploit its strengths while still making it possible for users' to configure it towards their needs.
- Bridging the abstraction gap for the translation between *hls4ml* high-level representation of neural networks and their customizable instantiation in hardware.

1.4 Contributions

The project aims to benefit the open-source community of ML researches that are in need of faster and more parametrizable neural network inference. The targeted audience for that operation are physicists at LHC, nonetheless, the hope is for the work to positively contribute in many ML fields by both offering a reliable tool for acceleration of existing designs and providing a useful resource for learning about the nature of reconfigurable hardware and its potential use for neural networks.

Chapter 2

Background

This chapter provides a closer look at the concepts required to understand this work. The following sections firstly discuss background and related work for topics in particle physics, then machine learning and finally reconfigurable hardware research.

2.1 Particle Physics

particle Collisions jets cern lhc level x trigger lhc Physics experiments are crucial Big data is crucial for Physics

2.2 Machine Learning

training-validation-test dataset training - inference - tensor - machine learning framework Transformer Neural Networks are great

2.3 Reconfigurable Hardware

c simulation, cosimulation, synthesis

roof ceiling model / pareto front latency resource usepackage throughput pipelining HLS hardware design (parallel vs serial etc) FPGA are great for NN FPGA are very hard-coded -> make the code deployable on any platform with optimal settings automatically HLS is difficult, so coding hardware in Python is desired -> make it easy for engineers and physicists to design systems Powerful hardware is king Metaprogramming allows for optimizations and customisability

Chapter 3

Project Plan

The aims of the project cover a wide range of challenges that form subsequent steps of accelerating neural networks while raising the abstraction layers and reducing domain-specific knowledge requirements. This naturally divides the work into smaller objectives that are described in details in the following paragraphs.

Firstly, the existing transformer neural network architecture has to be redesigned to accommodate for easier adaptation to non-general-purpose hardware. This comprises splitting layers into more basic components that are easier to map to hardware and abstract about as well as introducing hooks that collect different information during training and inference passes (e.g. running mean and variance for normalization layers, tensor sizes and values). At this phase some design choices are highlighted for further inspection where simplification or improvements can be made to greatly reduce the complexity and resource usage without crippling performance.

With the adapted software implementation, the next step involves recreating the architecture in HLS. Building the initial prototype tackles the difficulties related to the underlying differences between software and hardware development and results in an accurate, yet not optimal design. From there, an iterative process begins with acceleration hypothesis firstly tested in the original software model to ensure satisfactory accuracy and then getting expressed in HLS to quantitatively measure the latency and throughput differences. That is expected to yield a highly performant solution to the initial problem that is tailored to the specific FPGA constraints.

In order to overcome the innate limitations of "hand-tuning" a solution to a problem that varies both in time and between applications, the final step of the project relies on meta-programming strategies that automatically adapt the solution according to users' criteria, available platforms and overall experiment's aim. The list of approaches that can be taken here is nearly endless, however two key areas have be designated - adjusting the model according to the existing hardware to exploit its strengths as well as allowing for more abstract representation of an architecture in a well-known machine learning framework.

As previously mentioned, some initially planned ideas have already been implemented. The distinction between these and a more detailed look at the specific project tasks can be seen in figure 3.1.

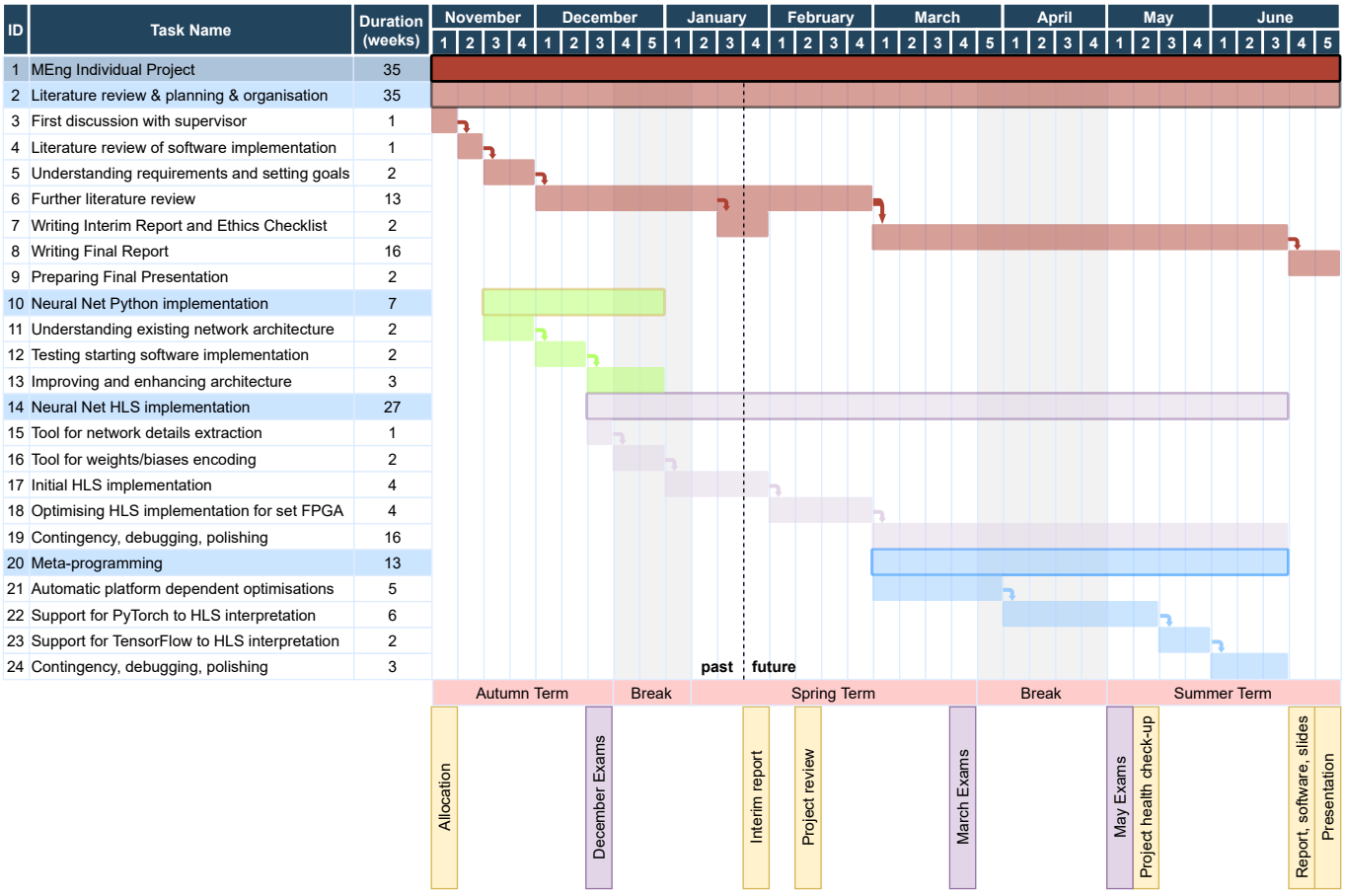

Figure 3.1: Project's Gantt chart representing initial plan of the work, past schedule has been updated to match ongoing progress accordingly

Chapter 4

Implementation

Chapter 5

Evaluation Plan

This section outlines the proposed evaluation plan for the project. The first objective of developing and optimizing a state-of-the-art neural network in hardware can be evaluated quantitatively, while integrating it into the *hls4ml* library and making it easy for new users to use requires a more qualitative approach.

5.1 Quantitative results

The following describes the quantities that are planned to be measured:

- Classification accuracy for each designed neural network on a validation dataset
- Inference latency and throughput when running on the target platform
- Hardware resource utilization (exact values for comparison with other platforms and percentage of available resources for understanding limitations):
  - Block RAM (BRAM)
  - Ultra RAM (URAM)
  - Digital Signal Processing units (DSP)
  - Flip-Flops (FF)
  - Look-Up Tables (LUT)

In the early stages of the project, the above quantities will be measured from the results from simulation and synthesis reports. At a later stage, the best designs will be run on actual hardware platforms to validate them under real-life use cases. The platform planned for this part is an Intel Stratix V FPGA hosted in a Maxeler MPC-X dataflow node with 8 Maia dataflow engines and 48 GB of DRAM.

Apart from clear design improvements, it is predicted that most evaluated designs will offer trade-offs between classification accuracy, inference throughput and hardware utilization. It is not possible to find a design that is superior in every way, hence a **Pareto front** will play a key role in understanding the overall performance and selecting configuration with specific needs in mind.

5.2 Qualitative results

Qualitative

Chapter 6

Ethical Considerations

The purpose of this project is to advance the next-generation particle physics experiments. There are two main aspects that need to be considered - the development of a hardware-mapped transformer neural network architecture and the easy-to-access translation and optimization toolchain for efficiently expressing networks in common machine learning frameworks.

The first feature is aimed at a purely civilian, scientific audience and it is tailored towards particle collision datasets. With that in mind, it is important to mention that, as with most machine learning research, there is potential for a misuse of the acceleration techniques towards a military or malevolent application that could negatively impact the society (issues A in table 6.1). However, this also means that there is a low risk for new emerging threats, rather the already present ones could become more serious. Fortunately, this should result in existing harm prevention measures to stay intact or solely require adjustments to their accuracy or speed thresholds.

With the second element's goal of making the creation and deployment of neural networks more accessible, it could be argued that this may in turn increase the number of physics experiments requiring high energy consumption, like those at LHC [9], thus negatively effecting the environment (issue B in table 6.1). However, this is considered a very low likely cause of action, as the research work of this project is aimed at helping already running experiments and more importantly, the negative environmental implications (for which there are various mitigation strategies [10, 11]) are heavily outweighed by potential beneficial technological advancements coming from the scientific discoveries.

Despite the aforementioned ethical issues, the project is aimed at benefitting the open-source scientific community world-wide. Its outcome could lead to a much more accessible and efficient inference methods that are applicable in many domains outside physics.

Table 6.1: Overview of potential categorized ethical issues with an indication of their applicability

	Involvement of...	Exists?
Humans	human participants	No
Personal data	personal data collection and/or processing	No
	collection and/or processing of sensitive personal data	No
	processing of genetic information	No
	tracking or observation of participants	No
	further processing of previously collected personal data	No
Animals	animals	No
Developing countries	developing countries	No
	low and/or lower-middle income countries	No
	putting the individuals taking part in the project at risk	No
Environment	elements that may cause harm to the environment, animals or plants	Yes (B)
	elements that may cause harm to humans	No
Dual use	potential for military applications	No
	strictly civilian application focus	Yes
	goods or information requiring export licenses	No
	affection of current standards in military ethics	No
Misuse	potential for malevolent/criminal/terrorist abuse	Yes (A)
	information on/or the use of sensitive materials and explosives	No
	technologies that could negatively impact human rights standards	Yes (A)
Legal	software for which there are copyright licensing implications	No
	information for which there are data protection or other legal implications	No
Other	any other ethics issues that should be taken into consideration	No

References

[1] CERN. Jets at CMS and the determination of their energy scale | CMS experiment, .

[2] Josh Cogan, Michael Kagan, Emanuel Strauss, and Ariel Schwartzman. Jet-images: computer vision inspired techniques for jet tagging. *The journal of high energy physics*, 2015(2):1–16, Feb 18, 2015. doi: 10.1007/JHEP02(2015)118. URL https://link.springer.com/article/10.1007/JHEP02(2015)118.

[3] Luke de Oliveira, Michael Kagan, Lester Mackey, Benjamin Nachman, and Ariel Schwartzman. Jet-images — deep learning edition. *The journal of high energy physics*, 2016(7):1–32, Jul 13, 2016. doi: 10.1007/JHEP07(2016)069. URL https://link.springer.com/article/10.1007/JHEP07(2016)069.

[4] Liam Moore, Karl Nordstrom, Sreedevi Varma, and Malcolm Fairbairn. Reports of my demise are greatly exaggerated: N-subjettiness taggers take on jet images. *SciPost physics*, 7(3):036, Sep 24, 2019. doi: 10.21468/SciPostPhys.7.3.036. URL https://hal.archives-ouvertes.fr/hal-01851157.

[5] Farah Fahim, Benjamin Hawks, Christian Herwig, James Hirschauer, Sergo Jindariani, Nhan Tran, Luca P. Carloni, Giuseppe Di Guglielmo, Philip Harris, Jeffrey Krupa, Dylan Rankin, Manuel Blanco Valentin, Josiah Hester, Yingyi Luo, John Mamish, Seda Orgrenci-Memik, Thea Aarrestad, Hamza Javed, Vladimir Loncar, Maurizio Pierini, Adrian Alan Pol, Sioni Summers, Javier Duarte, Scott Hauck, Shih-Chieh Hsu, Jennifer Ngadiuba, Mia Liu, Duc Hoang, Edward Kreinar, and Zhenbin Wu. hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices. Mar 9, 2021. URL https://arxiv.org/abs/2103.05579.

[6] Harvey B. Newman, Avikar Periwal, Maria Spiropulu, Javier M. Duarte, Maurizio Pierini, Eric A. Moreno, Aidana Serikova, Olmo Cerri, Jean-Roch Vlimant, and Thong Q. Nguyen. JEDI-net: a jet identification algorithm based on interaction networks. *The European physical journal. C, Particles and fields*, 80(1):1–15, Aug 14, 2019. doi: 10.1140/epjc/s10052-020-7608-4. URL http://cds.cern.ch/record/2688535.

[7] Abdelrahman Elabd, Vesal Razavimaleki, Shi-Yu Huang, Javier Duarte, Markus Atkinson, Gage DeZoort, Peter Elmer, Jin-Xuan Hu, Shih-Chieh Hsu, Bo-Cheng Lai, Mark Neubauer, Isobel Ojalvo, and Savannah Thais. Graph neural networks for charged particle tracking on FPGas. Dec 3, 2021. URL https://arxiv.org/abs/2112.02048.

[8] Xinyang Yuan. *ConstituentNet: Learn to Solve Jet Tagging through Attention*. PhD thesis, -09-22 2021.

[9] CERN. Facts and figures about the LHC | CERN, . URL https://home.cern/resources/faqs/facts-and-figures-about-lhc.

[10] R. Guida, M. Capeans, and B. Mandelli. Characterization of RPC operation with new environmental friendly mixtures for LHC application and beyond. *Journal of Instrumentation*, 11(07):C07016–C07016, jul 2016. doi: 10.1088/1748-0221/11/07/c07016. URL https://doi.org/10.1088/1748-0221/11/07/c07016.

[11] M. Capeans, R. Guida, and B. Mandelli. Strategies for reducing the environmental impact of gaseous detector operation at the CERN LHC experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated*

Equipment, 845:253–256, 2017. doi: https://doi.org/10.1016/j.nima.2016.04.067. URL https://www.sciencedirect.com/science/article/pii/S0168900216302807. ID: 271580.