

Imperial College London

MENG INDIVIDUAL PROJECT

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

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

Author:
Filip Wojcicki

Supervisor:
Prof. Wayne Luk

Second Marker:
Prof. Alexander Tapper

June 22, 2022

Abstract

Particle Physics studies the fundamental forces and elementary particles building the Universe. In order to verify the correctness of the theories, countless experiments have to be designed and carefully executed, with the main driving force of the myriads of engineers, physicists and researchers at the Large Hadron Collider (LHC) operated by the European Organization for Nuclear Research (CERN). With the unprecedented experiments' scale comes the challenge of accurate, ultra-low latency decision-making. Transformer Neural Networks (TNN) have been proven to accomplish cutting-edge accuracy in various domains, including classification for jet tagging, which is the target of this project. However, software-centered solution implemented for CPUs and GPUs lack the inference speed needed for real-time particle triggers.

This report proposes two novel TNN-based architectures efficiently mapped to Field-Programmable Gate Arrays (FPGAs). The first one outperforms the current state-of-the-art models' GPU inference capabilities by roughly 1000 times while maintaining comparable classification accuracy. The second one trades off some of its speed for accuracy and undergoes a broad design-space exploration which involves both pre-training and post-training quantization. The latter one leverages a custom-developed tool chain that augments existing solutions in terms of granularity and ease of use while following an innovative algorithm for relatively quick convergence.

In this project, several recently researched neural network components are designed to target FPGAs using High-Level Synthesis (HLS). The resulting open-sourced building blocks are both highly customizable and abstract, and aim to bridge the gap between hardware and software development, effectively reducing the time and complexity needed for creating efficient neural network hardware accelerators.

Acknowledgements

I would like to express my gratitude to Professor Wayne Luk for his guidance, insightful suggestions and constant encouragement throughout the project.

I would like to thank Professor Tapper for giving me a different view on the project's meaning and providing with the behind-the-scenes information about the LHC.

I want to thank Zhiqiang Que for his continuous technical support, our weekly meetings and always being available to answer any of my questions.

Lastly, I am very grateful for my family and friends whose support was invaluable during this project and the degree as a whole.

Contents

List of Figures	4
List of Tables	5
List of Listings	6
List of Algorithms	6
1 Introduction	7
1.1 Motivation	7
1.2 Objectives and Challenges	7
1.3 Contributions and Publication	8
1.4 Report Outline	8
2 Background and related work	9
2.1 Particle Physics	9
2.1.1 Standard Model	9
2.1.2 Particle Accelerators and Triggers at LHC	10
2.1.3 Dataset and Notations	10
2.2 Machine Learning	13
2.2.1 Metrics	13
2.2.2 Deep Neural Networks	13
2.2.3 Batch and Layer Normalization	14
2.2.4 Transformer Neural Networks and Self-Attention	15
2.3 Reconfigurable Hardware	16
2.3.1 Landscape of Hardware for Computing	16
2.3.2 High-Level Synthesis	17
2.3.3 hls4ml Codesign Workflow	18
2.3.4 Latency, Throughput, and Hardware Resources	18
2.3.5 Serial, Parallel, and Pipelined Architectures	19
2.3.6 Pareto Front and Roofline Model	20
2.4 Ethical Considerations	21
3 Architecture Exploration	22
3.1 Base Architecture	22
3.1.1 Input embedding and Residual Connections	23
3.1.2 Input Encoding	23
3.1.3 Normalization and Parameter Extraction	23
3.2 Ultra-Low Latency Architecture	24
3.2.1 Optimization and Tuning	24
3.2.2 Summary	25

3.3	Accuracy-Focused Architecture	25
3.3.1	Optimization and Tuning	25
3.3.2	Summary	26
3.4	Quantization-Aware Training	27
3.4.1	Floating-point and Fixed-point Representations	28
3.4.2	Existing Implementations	28
4	Hardware-Aware Analysis	31
4.1	Hardware-Aware Optimizations	31
4.1.1	Tensor Multiplication and Scaling	31
4.1.2	Softmax and Log Softmax Activations	33
4.2	Neural Networks Implementation	34
4.3	Analytical Latency and Resource Models	34
4.3.1	Latency Model	34
4.3.2	DSP Model	36
4.4	Post-training Quantization	37
4.4.1	Motivation	37
4.4.2	Constraints	37
4.4.3	Steps	38
4.5	hls4ml Contributions	39
4.5.1	Activation Functions	39
4.5.2	Normalization Layers	39
4.5.3	General-Purpose Blocks	39
4.5.4	High-Level Components	41
5	Evaluation	42
5.1	Proposed Models' Performance	42
5.1.1	Classification Accuracy	42
5.1.2	Receiver Operating Characteristic Curves	42
5.1.3	CPU and GPU Latency	44
5.1.4	Comparison with Existing Solutions	45
5.2	Hardware Acceleration	45
5.2.1	Ultra-Low Latency Model	46
5.2.2	Verifying Analytical Latency and DSP Models	49
5.2.3	Accuracy-Focused Model	49
5.3	Quantization Results	50
5.3.1	Quantization-Aware Training	51
5.3.2	Post-Training Quantization	52
6	Conclusion	54
6.1	Achievements	54
6.2	Limitations	54
6.3	Future Work	55
6.3.1	Alternative Transformer Design	55
6.3.2	Post-Training Quantization Advancements	56
6.3.3	Automatic High-Level-Synthesis Optimizations	56
6.3.4	Other Datasets	56
	References	57

List of Figures

2.1	Representation of different decay processes, based on the number of resulting jet clusters.	10
2.2	Representation of the distributions of feature values in the HLF dataset.	12
2.3	Diagram of a fully connected layer.	14
2.4	Left: diagram of a self-attention head. Right: illustration of H self-attention heads forming a multi-headed self-attention block.	15
2.5	From left to right: visualizations, for an input word, of the words focused by 1, 2 and 8 attention heads.	16
2.6	Diagram comparing serial and parallel configurations as well as showcasing designs with and without pipelining.	19
2.7	Example graph with designs plotted against quantities A/B, Pareto front highlighted.	20
2.8	Example graph with computational and memory bandwidth limitations showcasing the Roofline model	21
3.1	Diagram with an overview of the baseline architecture.	22
3.2	Diagram with an overview of the ultra-low latency architecture.	25
3.3	Simple illustration for equivalency between matrix multiplication using merged and unmerged weights on the example of <i>queries</i> , <i>keys</i> , and <i>values</i> used in self-attention.	26
3.4	Diagram with an overview of the accuracy-focused architecture; the inclusion of batch normalization before softmax within the multi-head self-attention block is not showcased for the sake of clarity.	27
3.5	Quantization noise from 32-bit floating-point to 3-bit (2-bit integer, 1-bit fractional) fixed-point representation; widths selected to highlight the behavior. Two methods of handling values outside representable range are showed: classic overflow as well as saturation to minimum or maximum value.	27
3.6	Comparison of a simple neural network prepared for quantization-aware training using Brevitas and QPyTorch frameworks.	29
4.1	Visualization of a tensor operation expressed in Einstein Summation notation.	31
4.2	Illustration of an adder tree with N inputs.	32
4.3	Direct hardware implementations of log softmax.	33
4.4	Optimized hardware implementations of log softmax.	33
4.5	Ultra-low latency architecture with highlighted pipeline stages. Not annotated blocks share stage with their neighbors.	35
4.6	3D plot of the latency model dependency on the number of transformer and feed-forward layers.	36
4.7	Visualization of a fixed-point number with its bit-widths as well as constraints imposed by a lookup table. For convenience, red and blue distinguish integer and fractional parts, while the darker hue shows the table-related parts.	38

4.8	Overview of the created implementations (dark-colored) and some existing components with similar functionality (light-colored).	39
5.1	Comparison between confusion matrices of the ultra-low latency model on the HLF dataset (left) and the accuracy-focused model on the constituent list dataset (right).	43
5.2	Comparison between ROC and AUC of the base (left) and ultra-low latency (right) models using the HLF dataset.	43
5.3	Comparison between ROC and AUC of the base (left) and accuracy-focused (right) models using the constituent list dataset.	44
5.4	Results of the optimization process for the ultra-low latency model.	47
5.5	Latency plotted against average resource utilization for the ultra-low latency model configurations.	48
5.6	Grid-search results - squares' area proportional to accuracy.	49
5.7	Illustration of the synthesis design space reaching the limit of available memory.	50
5.8	Performance against epochs for floating-point and fixed-point models.	51
5.9	Visualization of the optimal fixed-point precision of the types used in the quantized accuracy-focused model.	52
6.1	Landscape of efficient transformer architectures.	55

List of Tables

3.1	Comparison of standard precision types for floating-point values. Note the difference in half precisions, where <i>Brain</i> offers same the exponent width as single precision, and <i>binary</i> priorities fractional precision.	28
5.1	Comparison between classification accuracy, average ROC metrics, parameter count, and FLOPS between the baseline and the proposed architectures.	43
5.2	AUC and true positive rates at 10 % and 1 % false positive rates for the ultra-low latency and accuracy-focused models.	44
5.3	Comparison of ultra-low latency model's inference times, with batch size of 128	45
5.4	Comparison of accuracy-focused model's inference times, with batch size of 128	45
5.5	Summary of networks' inference time, accuracy, Floating-Point Operations Per Second and parameter number for optimal batch sizes, with best values in bold. Inference times were measured using [*] NVIDIA GTX 1080, [†] 16-thread virtualized cloud CPU, [‡] NVIDIA GTX 1080 Ti.	46
5.6	FPGA resources utilization for the ultra-low latency model.	47
5.7	Information about design's clock, latency and initiation interval for the ultra-low latency model.	48
5.8	FPGA resources utilization for accuracy-focused model.	50
5.9	Information about design's clock, latency and initiation interval for the accuracy-focused model.	50

List of Listings

4.1	From PyTorch <code>out = torch.einsum("qhc,khc->hqk", [A, B])</code> to HLS C++ code.	32
-----	--	----

List of Algorithms

3.1	Mechanism behind model parameter extraction	24
4.2	Algorithm for performing post-training quantization search	40

Chapter 1

Introduction

1.1 Motivation

LHC is the world’s highest-energy particle collider that is capable of producing and detecting the heaviest types of particles that emerge from events such as 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 collimated sprays of composite particles (hadrons) that are referred to as jets [1]. The initial particles created upon collision and their behaviors are of main interest of physicists, leading to jet tagging - the challenge of associating particle jets with their origin.

There are many detector types used for the analysis of the particle collisions, each based on a different physical phenomenon, which results in availability of both higher and lower level features from each event. 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 trained on higher-level data achieving nearly identical accuracy as dense neural networks trained on the data from the other end of the spectrum [4].

The information throughput of Petabytes per second 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 of utmost interest, hence this paper is motivated by this challenge which includes exploring various types of neural network architecture as well as the necessary infrastructure and deployment processes. Recently, *hls4ml* codesign workflow have been successfully adopted in particle physics experiments [5], which 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’ inference capabilities. 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 [6, 7] and transformer neural networks [8].

1.2 Objectives and Challenges

The purpose of this project is to develop novel, hardware-aware neural network architectures as well as to establish efficient ways mapping them onto FPGAs. Another objective is to use metaprogramming strategies to integrate them into the *hls4ml* library or standalone tools, with various optimizations approaches that offer trade-offs between latency, throughput and hardware resources usage. Hence, there is an emphasis on creating parametrizable and reusable designs that can support creation of ultra-low latency systems, effectively transforming proof-of-concept implementations into optimized hardware accelerators.

The two main challenges of the project involve:

- **C1:** Developing deep and complex neural network models with inference times in the microseconds range, which requires much lower abstraction levels than a typical ML framework. It is also crucial to stay aware of the underlying hardware architecture to exploit its strengths while keeping compile-time and run-time configuration easily accessible.
- **C2:** Bridging the abstraction gap for the translation between high-level representation of neural networks and their optimized mapping to hardware. The design space exploration is a long and difficult process, which needs careful examination and analytical performance models to find the optimal solutions.

1.3 Contributions and Publication

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

The novel aspects proposed in this work can be summarized as follows, with the first two points addressing challenge **C1**, and the other two points challenge **C2**:

- **Hardware-Aware Design:** Two novel transformer architectures are designed with hardware in mind, exploiting the target hardware platform to achieve cutting-edge latency and accuracy.
- **Efficient Hardware Implementation:** hls4ml codesign workflow is extended with a number of complex neural network layers that benefit from a range of optimization techniques.
- **Quantization-Aware Training:** Several available solutions are surveyed and described, and the best one in terms of applicability to state-of-the-art neural network designs is selected and adapted to transformer neural networks.
- **Post-Training Quantization Search:** A custom algorithm is proposed that finds the optimal set of bit-widths for a High-Level Synthesis model according to user-defined constraints.

The bulk of the work and analysis conducted in this project were summarized in the paper "Accelerating Transformer Neural Networks on FPGAs for High Energy Physics Experiments" and submitted in the long paper category to the 18th *International Symposium on Applied Reconfigurable Computing*. A journal article derived from this project is being prepared for publication.

1.4 Report Outline

This report begins by discussing the necessary particle physics background to understand the scope of the work, followed by the related work in the field of machine learning, with an emphasis on the state-of-the-art architectures, and a deeper dive into the reconfigurable hardware technology in chapter 2. Then in chapter 3, the two proposed novel neural network architectures are described in details, including the necessary training and processing steps as well the quantization-aware training. After that, chapter 4 covers the hardware implementation of the architectures, provides analytical models for latency and hardware usage, and described the novel post-training quantization algorithm. Chapter 5 discusses the evaluation metrics and collected experimental results, which is concluded by chapter 6, which also proposes future work that extends upon this analysis.

Chapter 2

Background and related work

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. At the end of this chapter, the ethical issues that could arise from the project are listed and discussed.

2.1 Particle Physics

To be able to understand the scope of the project and the applicability of the work in modern research, this chapter gives an overview of the key concepts from particle physics that appear through the paper. The explanation is written for readers with no prior background in physics.

2.1.1 Standard Model

The Standard Model is a theory that describes the connections between weak, strong and electromagnetic interactions, which are three of the fundamental forces. The possible unification with the forth one - gravity - is an ongoing research [9], and while certainly outside the scope of this project, it should be noted that some of the physical experiments that this work explores aim to help with it [10, 11].

The Standard Model also provides a classification of all the elementary particles. A non-exhaustive list of them is described below, with particles that this report is concerned about (as they appear in the proton-proton collisions) being highlighted.

- Fermions
 - Leptons - participate in electroweak interactions; include electron (e^-)
 - Quarks - participate in strong interactions; include **light quarks** (q)¹ and **top** (t) **quark**
- Bosons
 - Gauge bosons - force carriers; include photon (γ), **W boson** (W^+ , W^-), **Z boson**, **gluons**
 - Scalar bosons - give rise to mass; include **Higgs boson** (H^0)

The information about the decay processes seen in equations 2.1, 2.2, 2.3 form the dataset of this report, with a visualization in figure 2.1 (obtained from [6]). It is important to note that where applicable, the particles on the left-hand side of the arrows undergo a series of decays before reaching the right-hand side, when the only particles left are those composed of quarks and antiquarks (denoted by the vertical bar), referred to as hadrons.

$$q/g \tag{2.1}$$

$$H^0/W/Z \rightarrow q\bar{q} \tag{2.2}$$

$$t \rightarrow Wq \rightarrow q\bar{q}q \tag{2.3}$$

¹Light flavor quarks: up (u), down (d), charm (c), and strange (s) quarks

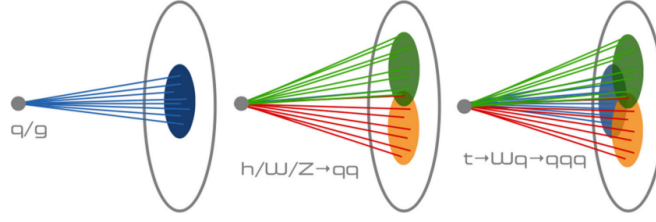


Figure 2.1: Representation of different decay processes, based on the number of resulting jet clusters.

2.1.2 Particle Accelerators and Triggers at LHC

The two LHC experiments that are of most concern in this report are CMS and ATLAS. They are both large general-purpose particle detectors, that were notably involved in the discovery of the Higgs boson [12]. Several processing steps happen between particles colliding and theories being proven, however real-time particle detectors comprise the very first elements of this pipeline. They are composed of triggers split into the following levels [13, p.16]:

- **Level 1 trigger (L1T)** - it is implemented in hardware (FPGAs) and firmware, it is pipelined (a term explained in details in subsection 2.3.5) and it cannot allow for any dead-time, which means that it has to continuously process data with a fixed latency.
- **Level 2 trigger (L2T)** - it is implemented in hardware and software and can include regional processing.
- **Level 3 trigger (L3T)** - it is implemented in software, using farms of CPUs. It is close in behavior to non-real-time algorithms.

LHC operates in intertwined periods of operation and shutdown. The latter come from the demanding nature of the experiments that necessitates maintenance and upgrades to the apparatus and machinery, as well as the science and engineering advancements which allow for more efficient algorithms and technologies to be adopted. Very recently, after four years of break, LHC restarted experiments, which marks the beginning of "Run 3" [14]. Since their origin, the L2T and L3T have been merged into High Level Trigger (HLT) [15, p.47], which is planned to rely on thousands of multithreaded CPUs and GPUs. As for the L1T key specifications that will be used to evaluate the design in this paper, its input data frequency is 40 MHz, which with a pipeline depth of 500 results in a $12.5 \mu s$ latency, and its output frequency to HLT is equal to 750 kHz.

2.1.3 Dataset and Notations

The datasets used in this work has been simulated to mimic the 13 TeV proton-proton collisions performed at LHC, and it includes information about the most energetic jets [16] (30 [17], 50 [18], 100 [19] and 150 [20]) that were constructed using the anti- K_t clustering algorithm [21]. A number of jet representations are available in the dataset:

- High level features (HLF), which are physically inspired,
- Images, which are related to an energy heat-map,
- Constituent list, which contains jets' constituent hadrons from the following list: light quarks, top quarks, W bosons, Z bosons, and gluons.

Compared to the other two, the constituent list is a lower-level representation, however, as mentioned in chapter 1, this should not affect the classification accuracy [4]. It is also worth mentioning that a simpler dataset that contains only the HLF jet representation [22] is also used in this project as it vastly reduces the complexity of a design while offering comparable accuracy. A more thorough discussion between the differences in their use cases is carried in chapter 5, but it is worth mentioning that the HLF representation has been successfully used in conjunction with deep neural networks [22], while the images and constituent lists were adopted for graph neural networks [6].

To facilitate further analysis, this subsection also explains the notation used for the dataset as it is important throughout the report. Each constituent element \mathbf{x}^l is a 16-dimensional vector, where l denotes the index in the list:

$$\mathbf{x}^l = [x_0^l \ x_1^l \ \dots \ x_{15}^l]^T \in \mathbb{R}^{16} \quad (2.4)$$

The physical meaning of each element's dimension is not taken into consideration, and all of them are treated as equally important. A constituent list \mathbf{x}^i acts a single sample, with index i within a dataset, and it varies in terms of the number of constituents L , but it has no more of them than the dataset name suggests:

$$\hat{\mathbf{x}}^i = [\mathbf{x}^{i,0} \ \mathbf{x}^{i,1} \ \dots \ \mathbf{x}^{i,L-1}] \in \mathbb{R}^{L \times 16} \quad (2.5)$$

Using the 30 jet dataset as an example, each sample has between 1 and 30 constituents, although in the majority of samples it is the upper boundary. Hence, the whole dataset with N samples can be represented as D :

$$D = \begin{bmatrix} \hat{\mathbf{x}}^0 \\ \hat{\mathbf{x}}^1 \\ \vdots \\ \hat{\mathbf{x}}^{N-1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{0,0} & \mathbf{x}^{0,1} & \dots & \mathbf{x}^{0,L-1} \\ \mathbf{x}^{1,0} & \mathbf{x}^{1,1} & \dots & \mathbf{x}^{1,L-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}^{N-1,0} & \mathbf{x}^{N-1,1} & \dots & \mathbf{x}^{N-1,L-1} \end{bmatrix} \in \mathbb{R}^{N \times L \times 16} \quad (2.6)$$

The main jet datasets contain 880,000 samples regardless of the number of jets per sample, while the HLF dataset contains 830,000 examples, which are split into training and test samples in 70 : 30 and 80 : 20 proportions accordingly. In both cases, the datasets are balanced, meaning that they contain all available particle classes in near identical proportions. This is a desired characteristic of data as it avoids having to introduce weighted results as a measure of protecting the model from not learning the underrepresented classes properly. When it comes to the distribution and the range of values for each feature, their visualizations can be seen in figure 2.2 and TODO for HLF and constituent list datasets respectively.

It is also worth mentioning that normalization² is the only preprocessing measure used in this work, and it is only applied to the HLF dataset. The reasons behind avoiding it in case of the constituent list dataset are discussed in detail in section 3.3. Both dataset samples are simulated and do not contain any illegal or null values that would require dropping or substituting. There is also no need for data augmentation because of the simulated origin of data.

Constituent list feature distribution

|

|

|

|

|

|

|

²Process of subtracting mean and dividing by standard deviation, also referred to as "standardization".

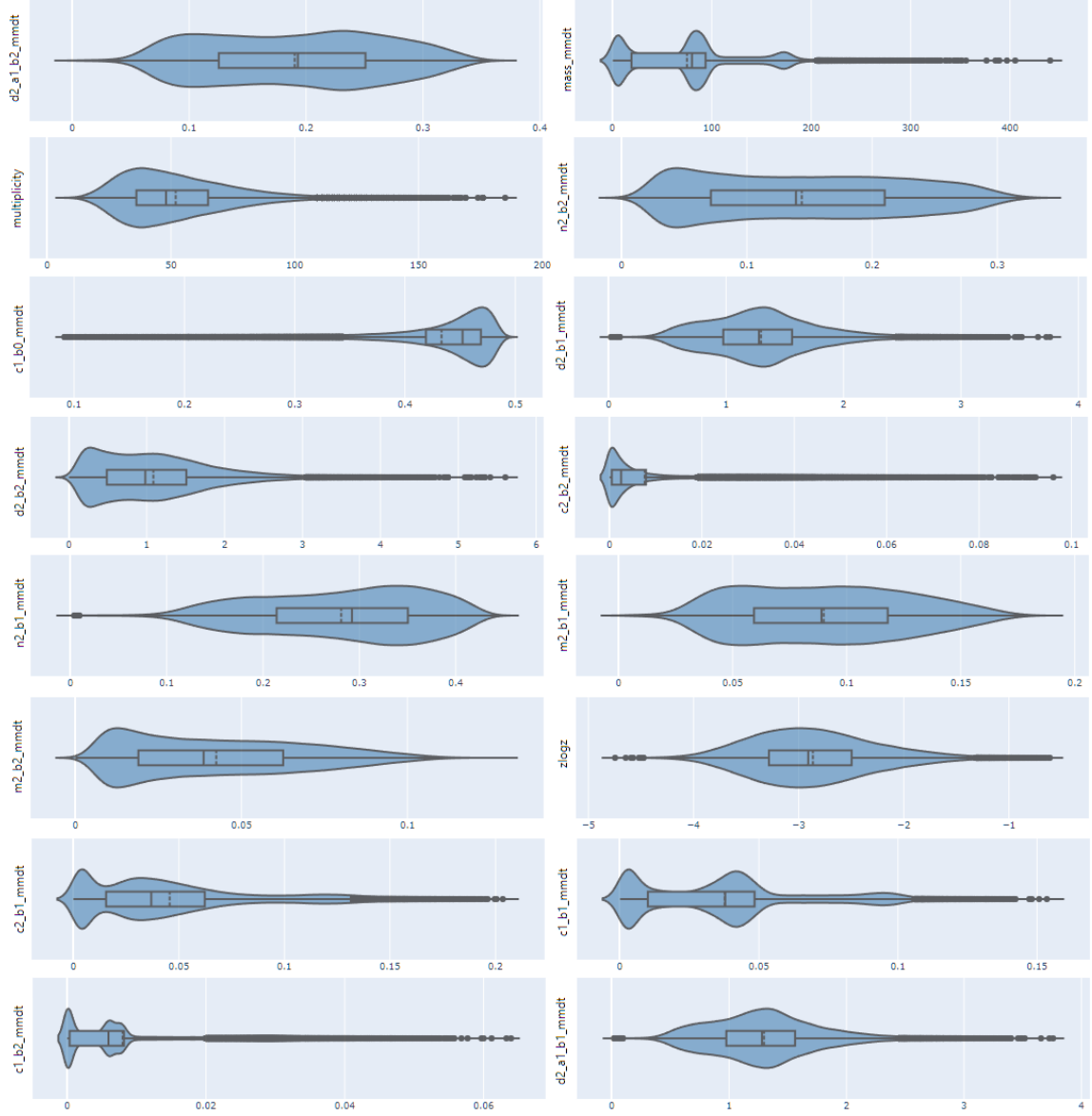


Figure 2.2: Representation of the distributions of feature values in the HLF dataset.

2.2 Machine Learning

Neural networks belong to a wider field of machine learning (ML) - the study of using experience to improve algorithms. This section assumes a basic understanding of ML and gives a brief overview of the topics needed to understand the scope of the project. It then explains in more details the background and related work for the architectures involved in this research.

2.2.1 Metrics

There are several key metrics used for assessing the success of an ML algorithm, and the following will be used throughout the report:

- **Classification accuracy** - a simple measure of the percentage of correctly classified samples.
- **Area Under the Curve (AUC) for the Receiver Operator Characteristic (ROC)** - a more complex measure of the model's ability to correctly distinguish between classes. It can be used similarly to the classification accuracy, but it favors discriminative over representative models.
- **Confusion matrix** - a tabular metric that compares the actual samples' classes with the predicted ones, effectively categorizing results into four groups: true positive, false positive, false negative, and true negative. This allows for an easy calculation of precision and recall values.

Let's consider an unbalanced dataset where a classifier always selects the most common class. This is clearly a poor choice of a classification method, however it is likely to yield high accuracy, obfuscating the actual classifier's performance. AUC considers the true positive rates across the whole range of false positive rates, which solves that problem while also offering a more comprehensive view of the results. Confusion matrix can be used for finding classes that get confused with each other, highlighting the areas which the classifier should be improved in.

2.2.2 Deep Neural Networks

While there exist a number of ML techniques that have proven successful for various use cases at LHC, like Support Vector Machines [23] or Boosted Decision Trees [24], in the last years deep neural networks (DNN) have been proposed with improved results for applications like infrastructure monitoring [25], offline data analysis [26], and the main interest of this report - detectors' trigger mechanisms.

In many use cases the neural networks architectures are optimized and accelerated to shorten the training time (measured in hours or even days) to reduce the time needed for evaluating different design configurations, and easily performing the hyperparameter search. However, this work focuses on accelerating the inference to match the extremely low latency required in the LHC detectors' L1 triggers. Although often measured in milliseconds, sub-milliseconds inference time has been achieved for this application with the use of FPGAs using architectures for basic DNN [22], and recently sub-microseconds latency for graph neural networks (GNN) [27, 28]. These implementations serve as a baseline latency for this project which aims to achieve comparable performance with higher AUC value.

Figure 2.3 shows an overview of a fully connected neural network with one hidden layer, which allows to derive the mathematical formulae. Each layer consists of neurons which hold a value, which means that input, output and the intermediate hidden layer can be modelled similarly. The arrows between neurons represent learnable weights, while the (optional) biases involved in the calculation, denoted as b , can be depicted as arrows from the "bias neurons", in all but the last layer. The value of a neuron depends on all the neurons in the previous layer as well as the weights and biases between them, which can be formulated into the equation 2.7 using the hidden layer as an example:

$$h_0 = f(w_{i,0} \cdot i_0 + w_{i,1} \cdot i_1 + \dots + w_{N-1} \cdot i_{N-1} + b_{i,0}) = f\left(\sum_{j=0}^{N-1} w_{i,j} \cdot i_j + b_{i,0}\right) \quad (2.7)$$

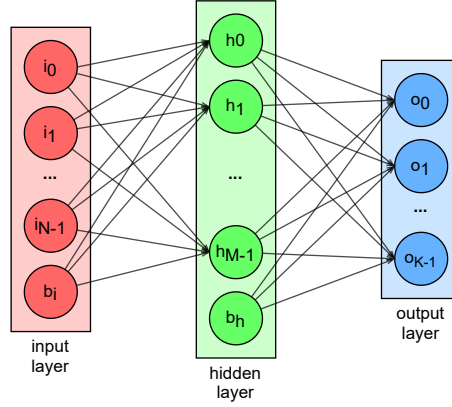


Figure 2.3: Diagram of a fully connected layer.

What is also not displayed in the diagram, but can be seen in the equation is the activation function f that is required to introduce non-linearity in the computations. Without it, all consecutive layers involving solely multiplication and addition could be simplified to a single layer thanks to the distributive property in linear algebra, defeating the point of having multiples of learnable parameters. The activation function can be as simple a piece wise linear function called Rectified Linear Unit (ReLU) defined in equation 2.8:

$$\text{ReLU}(x) = \max(x, 0) = \begin{cases} x & x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.8)$$

or more complicated like the Sigmoid Linear Unit (SiLU), based on the sigmoid function, both presented in equation 2.9

$$\text{SiLU}(x) = x \cdot \sigma(x) = x \cdot \frac{1}{1 + e^{-x}} \quad (2.9)$$

As layers are tightly connected to each other, this type of neural layer is often referred to as *fully connected* or *linear*. It requires a relatively large number of separate weights and biases, which makes it both computationally and memory intensive, but nonetheless, modern network architectures can have dozens of these layers, not to mention plethora of other types.

Since the invention of deep neural networks, countless other network classes have been developed. The list below gives a brief explanation of those that have been experimented with for jet-tagging.

- **Convolutional Neural Networks (CNNs):** TODO
- **Graph Neural Networks (GNNs):**
- **Recurrent Neural Networks (RNNs):** GRU, LSTM

2.2.3 Batch and Layer Normalization

Batch norm vs layer norm as both are used in the architecture

|

|

|

|

2.2.4 Transformer Neural Networks and Self-Attention

A promising architecture that has been chosen as the topic of this project is the transformer neural network (TNN). Similarly to RNNs, TNNs were designed for sequential input data, most commonly found in natural language processing applications, however, compared to RNNs, they process all input data at once. In RNNs, convolutional [29] or attention mechanisms [30] are used in a recurrent manner to allow models to learn the representation and connections between different parts of the input sequence, which most commonly are words in a sentence. This limits the parallelizability as the network is handled serially - each hidden state needs to wait for the result generated by the previous one. In TNNs, a modified mechanism called self-attention [31] is used which can find global relations in a data, without relying on the temporal, sequencing information. The self-attention combines several simpler operations to achieve its strength, including linear layers, matrix multiplication and the softmax function, formula for which is presented in equation 4.4.

$$\text{softmax}(x) = \frac{\exp(x_i)}{\sum \exp(x_i)} \quad (2.10)$$

Softmax can be described as mapping a vector to a ratio between each input's exponentiation result and the sum of all such values, which gives the property of the resulting vector entries sum equal to one. This characteristic means that the output can be treated as vector of probabilities, which is often exploited in the final activation layer of a neural network.

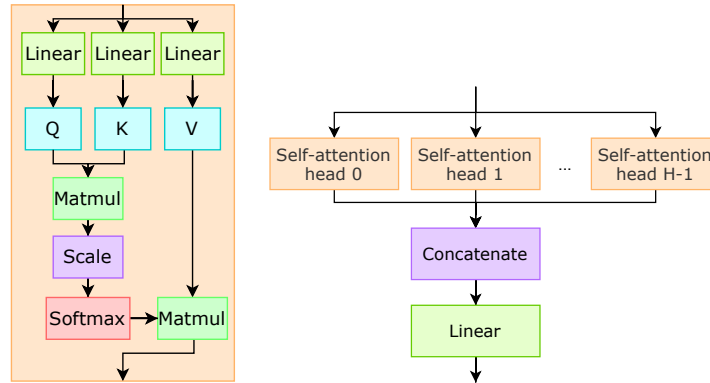


Figure 2.4: Left: diagram of a self-attention head. Right: illustration of H self-attention heads forming a multi-headed self-attention block.

A diagram representing the initial implementation of the self-attention can be seen on the left in figure 2.4. The Q , K , and V stand for *queries*, *keys*, and *values* respectively, which although arbitrary, are meant to give a better understanding behind the idea of this mechanism. It is also important to note that multiple blocks of self-attention, referred to as *heads*, can be used together, which allows for each head attending information about a different hidden characteristic of an input. The results of all heads are simply concatenated, increasing output's dimensionality, and multiplied with a learned weight, as seen on the right in figure 2.4. To better comprehend the interactions between information learned by the heads, figure 2.5 shows a visualization for 1, 2 and 8 heads on an example sentence, obtained using Tensor2Tensor library [32, 33]. While it is quite clear in the example that "it" is mostly associated with "The" and "animal" with 1 attention head, the interpretability is worse in case of 2 and 8 attention heads.

The complexity of a multi-head self-attention depends on the dimensions of the underlying *queries* $\in \mathbb{R}^{N_Q \times d_Q}$, *keys* $\in \mathbb{R}^{N_K \times d_K}$, and *values* $\in \mathbb{R}^{N_V \times d_V}$ parameters. The resulting time complexity is represented in the equation 2.11, while the space complexity can be seen in equation 2.12 with h and d_{out} standing for number of heads and output dimensions respectively in both cases.

Cite or derive this complexities

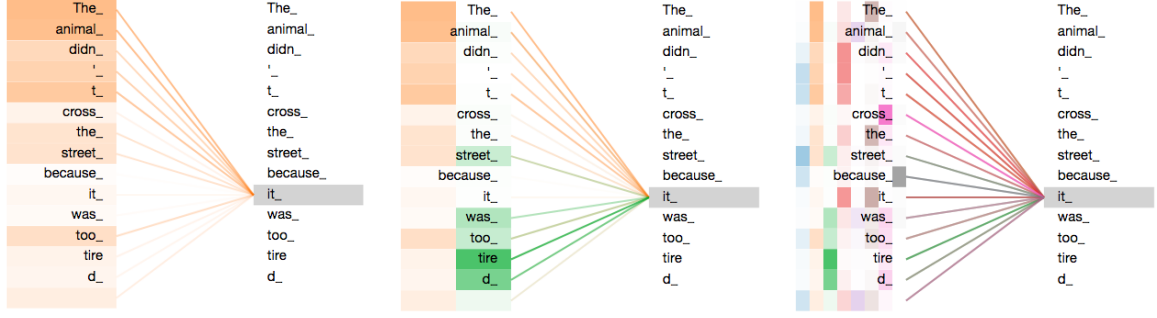


Figure 2.5: From left to right: visualizations, for an input word, of the words focused by 1, 2 and 8 attention heads.

$$\mathcal{O}(h \cdot (N_Q N_K (d_Q + d_V) + d_Q^2 (N_Q + N_K) + d_V^2 N_K + N_Q d_V d_{out}))) \quad (2.11)$$

$$\mathcal{O}(h \cdot (N_Q (N_K + d_V) + d_Q (N_Q + N_K) + d_V N_K + N_Q d_{out}))) \quad (2.12)$$

It has to be mentioned that in terms of the AUC value, a recent implementation of a transformer called ConstituentNet [8] has been shown to outperform previous state-of-the-art GNN implementations like JEDI-net [6] and thus serves as an inspiration for the starting point architecture of chapter 3, which is entirely devoted to a further analysis of the network design and suitability for jet tagging. More specifically, given the strong results of its software implementation, an FPGA-mapped design is believed to have a possibility to be a viable alternative to existing designs for L1T at LHC.

2.3 Reconfigurable Hardware

A significant portion of the project's work involves exploiting reconfigurable hardware to vastly reduce the inference time of the proposed neural networks. This section explains in more detail the technology and characteristics of reconfigurable hardware, and in particular, FPGAs.

2.3.1 Landscape of Hardware for Computing

The modern landscape of digital integrated circuits (IC) is very rich can be divided into numerous categories depending on the technology used and expected functionality [34]. A list of platform types is described below, with the emphasis of their suitability for neural networks applications.

- **Central Processing Units (CPU)** - the most commonly found ICs that are at the core of personal computers, laptops and handheld devices. They are capable of executing a broad range of predefined instructions. As CPUs have become widely adopted in research long before the emergence of the other technologies from this list, they were the first platforms for the training and inference of neural networks with promising results back in the 1980s and 1990s for applications like high energy physics [35] or biology [36]. Although it is possible to achieve speed-ups of over 10x the baseline performance with careful optimizations [37], CPUs are now consistently outperformed by more suitable technologies, and only limited to certain inference tasks.
- **Graphic Processors (GPU)** - ICs originally specialized in graphics processing intended for displaying images. Since their inception, due to the type of calculations involving matrix and vector operations, other applications related to cryptography and neural networks have also adopted GPUs as their main platform. In the former domain, cryptocurrency mining has transitioned from CPU to GPU to increase profitability [38], while for the latter, the more powerful hardware drastically reduced training and inference times, thus allowing for deeper and more complex architectures yielding higher accuracy [39, 40].

- **Application Specific Integrated Circuits (ASIC)** - as suggested by the name, those are the custom designed ICs heavily specialized for a particular application. It is hard to generalize them, as the use cases can cover any modern computing problem, but the commonality is a vast improvement in performance and power usage compared to more general purpose solutions. However, the long and expensive development process pose extremely high barriers to entry for most users. Fortunately, off-the-shelf products like the Graphcore Intelligence Processing Units [41], that are designed specifically with machine learning applications in mind, as well as other custom designs [42, 43] are starting to offer a compelling platform for working with neural networks.
- **Field-Programmable Gate Arrays (FPGA)** - differently from the previous listed IC types, FPGAs are not manufactured for a specific use case, and in fact, they can be reprogrammed to be a platform for a different application at any time. The reprogrammability comes at a cost of performance and power consumption compared to ASICs [44], but at the same time outperforms GPUs in these regards [45, 46]. It is also suggested, that with some technological improvements focused on ML applications, FPGAs can narrow the gap between ASICs without needing to stick to one particular design [44, 47, 48].

FPGAs offer an interesting trade-off between implementation effort and acceleration potential when it comes to neural networks and for that reason they have been chosen as the target technology in this report. The following subsections give a closer look at some of their characteristics and associated tools.

FPGA lattice overview to visualize explain the idea behind this technology

|

|

|

2.3.2 High-Level Synthesis

For many years, FPGAs have been modelled using register-transfer level (RTL) design abstraction with the use of hardware description languages like Verilog or VHDL. However, to increase productivity and allow for a more convenient design state space exploration (DSE), a more abstract modelling process called High-Level Synthesis (HLS) can be adopted. The design can be expressed in a software programming language like C, C++ or Java, which can be both manually and automatically optimized, and transformed to an equivalent RTL. This is especially beneficial in research, where compared to industrial environment, it is more likely that a slightly lower quality of results can be afforded for increased productivity and easier DSE. In fact, a recent study shows that on average only one third of design time and half of the lines of code are needed for an equivalent project done in HLS in comparison to RTL while the quality of results varies and can even outperform the RTL implementations for some applications [49].

This report's work is based on Xilinx Vivado HLS design suite. When developing a solution, it is important to note, that the synthesis process can take a significant amount of time (from a couple of hours to days on a modern powerful machine), and so there exist two simulation methods - a C-simulation that can quickly and directly evaluate a software benchmark against an emulation of the design, and a more truthful, cosimulation that firstly synthesizes a design and accompanying test bench to RTL and then performs an RTL simulation. A final, definitive evaluation of the results requires programming a target FPGA with the generated bit stream of the design and exchanging input/output data with a program that usually runs on a CPU.

HLS to RTL flow diagram

|

|

2.3.3 hls4ml Codesign Workflow

A commonality between the recent best performing hardware-mapped neural network models is the use of the *hls4ml* codesign workflow that was mentioned in section 1.1.

More about hls4ml

Difficulty: rtl > hls > python hl4ml, draw comparison with assembly

|

|

|

|

2.3.4 Latency, Throughput, and Hardware Resources

To properly navigate during the DSE and assign scores the solutions, the following characteristics have to be considered:

- **Latency** - A time measure of a system between receiving an input signal and producing a *corresponding* output. It is crucial in real-time processing where it has to be lower than the period between subsequent input samples. Depending on the application, latency in the microseconds or nanoseconds range can be expected from an FPGA. To recall from section 2.1.2, the latency constraint for this work comes from the specification of L1T at LHC and is equal to $12.5 \mu s$.
- **Throughput** - A rate of samples processed in a unit of time. For architectures that only start to process new elements after the previous one has finished, it is directly linked to latency. However, in modern ICs, especially in FPGAs, it is one of the defining metrics of performance and designs tend to exploit pipelining and parallelizability to marginally trade off their latency to increase it. Despite that, in this work, this measure is of little interest given the fixed latency and no dead-time constraint of L1T.
- **Resource utilization** - A more complicated, often multidimensional, metric that describes either the raw number or ratio of total usage of the hardware components of an FPGA. Typically, the higher it becomes, the more power is drawn by an FPGA, however, it is most often used to guide the design process to avoid running out of a certain resource. This can be done by potentially deploying an alternative method that can be implemented using a different, less contested resource.

While FPGAs vary in terms of hardware resource configurations, several key components can be distinguished:

- **Block Random Access Memory (BRAM)** - RAM unit which offers higher capacity storage with a relatively high access latency. It is best for storing big chunks of data that are not continuously used.
- **Lookup Table (LUT)** - a simple memory component that can store a very limited number of values while offering low access latency. It is often used for storing precomputed function values which can remove complex calculations from the critical path.
- **Flip-Flop (FF)** - more precisely, a D flip-flop, which implements a basic 1-bit binary register.
- **Digital Signal Processing (DSP) slice** - a logic element that is very efficient at additions, subtractions, multiplications, and boolean logic operations. It can process more than two inputs at once, which efficiently handles common algebraic operations like multiply-accumulate.

To fully understand the trade-offs between designs, one cannot forget about the metric directly related to the specific task that is accelerated in hardware. In the case of this report, classification accuracy and AUC described in section 2.2.1, will also play key roles in evaluating various configurations.

2.3.5 Serial, Parallel, and Pipelined Architectures

Hardware architectures use components that can be configured in different ways depending on the overall goal or a limiting factor. The high-level configurations of building blocks are displayed in figure 2.6 and can be described as follows:

- **Serial** - elements are arranged in a chain, processing one after another. This way uses less resources than an equivalent parallel configuration by reusing given components R times, thus approximately trading-off R times less required resources for R times higher latency.
- **Parallel** - elements share a common input and start processing data at the same time. This way ends in a lower latency than an equivalent serial configuration consuming using more resources.
- **Pipelined** - a more sophisticated arrangement, in which subsequent processing blocks (that can be either placed serially or in parallel) form a pipeline of processing stations separated by simple storage elements called pipeline registers, implemented using FFs. This maximizes the usage of the design blocks, hence increasing throughput with a minimal sacrifice of latency and resource usage.

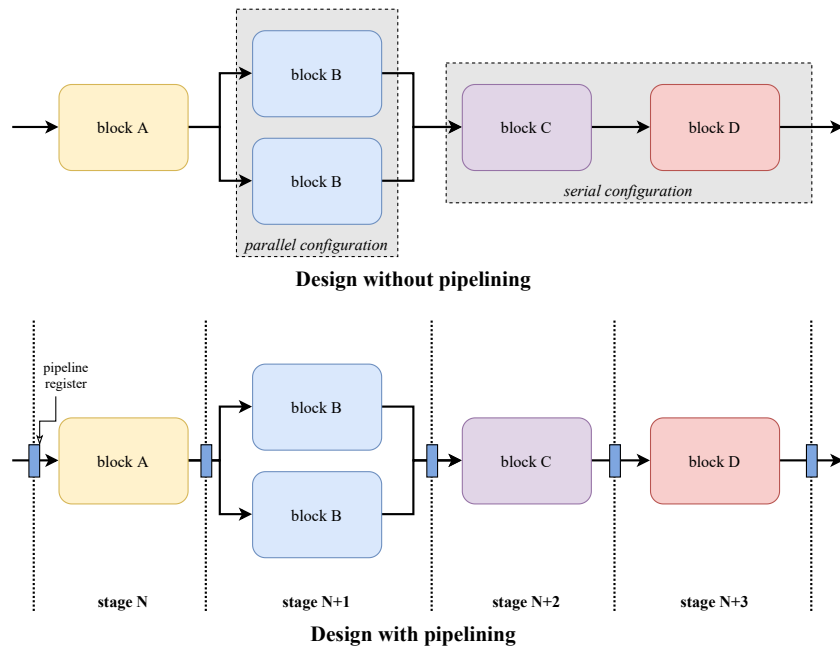


Figure 2.6: Diagram comparing serial and parallel configurations as well as showcasing designs with and without pipelining.

2.3.6 Pareto Front and Roofline Model

To make an informed design decision, various architectures can be compared by arranging them on a dependency graph (e.g. latency vs resource usage) and observing the Pareto front - the set of solutions for which there are no better ones in regard to one quality given that the other measure is not worse. The slightly complex definition can be easily understood from figure 2.7, which also highlights another use of this method - finding design configurations that are yet to be explored.

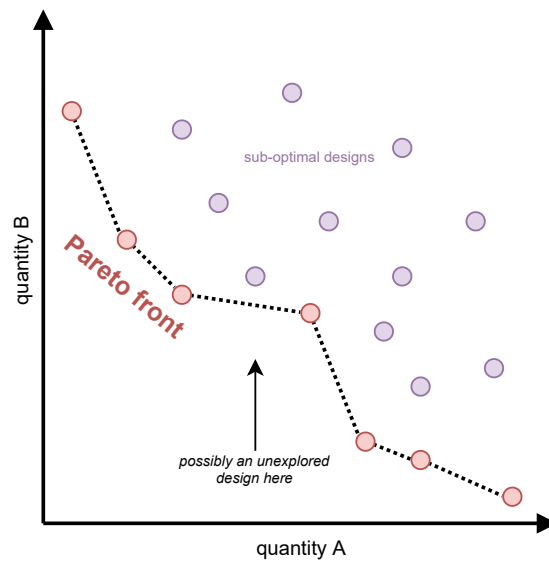


Figure 2.7: Example graph with designs plotted against quantities A/B, Pareto front highlighted.

Another intuitive performance visualization comes in the form of the Roofline model, which compares the obtained results with theoretical limits coming from inherent hardware limitations like clock frequency or memory bandwidth. An example can be seen in fig 2.8.

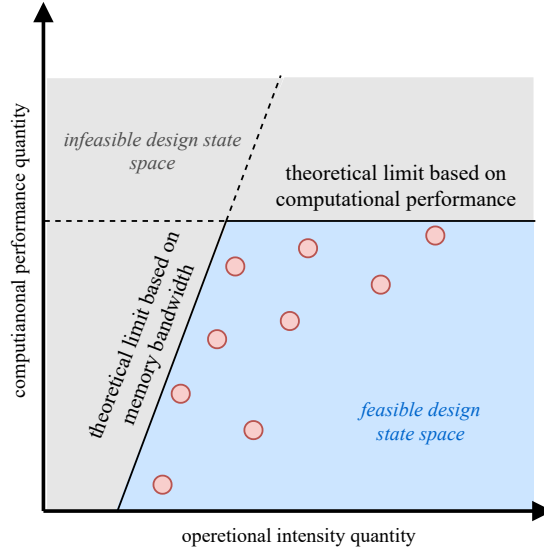


Figure 2.8: Example graph with computational and memory bandwidth limitations showcasing the Roofline model

2.4 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 and 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. However, this also means that there is a low risk for new emerging threats from this particular work; rather the already present ones could become slightly more serious. Fortunately, this should result in existing harm prevention measures staying intact or solely requiring 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 high energy physics experiments requiring immense energy consumption, like those at LHC [50], thus negatively affecting the environment. 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 [51, 52]) are heavily outweighed by potential beneficial technological advancements coming from the scientific discoveries.

Aside from 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 particle physics.

Chapter 3

Architecture Exploration

This chapter presents the proposed neural network architectures. It starts with a baseline TNN network implemented in PyTorch, which then undergoes a series of hardware-aware adaptations specific to jet tagging. During this process two separate architectures are developed, which differ by the input type and design goal. The first one, referred to as the *ultra-low latency* one, targets the HLF jet representation and aims to achieve the lowest possible latency at the cost of accuracy and AUC values. The second one, called *accuracy-focused*, is based on the constituent list jet representation and trade-offs latency for quality of classification while still remaining within L1T timing constraints. In the end, the mechanism of quantization-aware training is explained, with a proposed way of implementing it in this project.

3.1 Base Architecture

The starting point of this analysis is derived from transformer architecture used in the original paper [31] and recent proof-of-concept used for jet tagging [8]. The overview of the network components can be seen in figure 3.1.

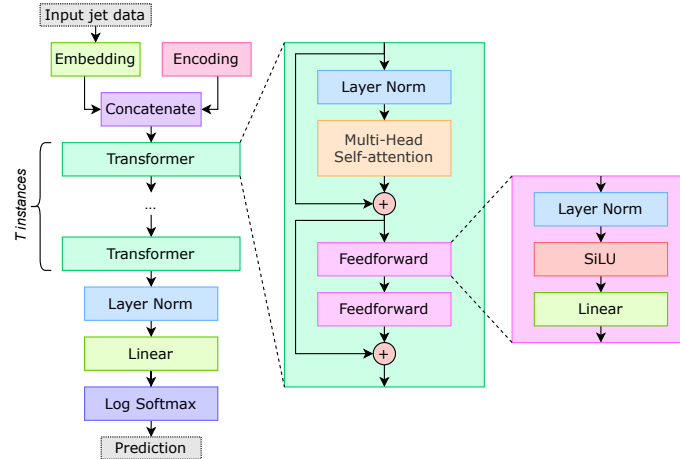


Figure 3.1: Diagram with an overview of the baseline architecture.

The straight-forward path between model's input and output highlights the sequential nature of transformer which stands in opposition to recurrency present in GRU and LSTM models. While this allows for the aforementioned parallelizability and pipelining on FPGAs, it also poses a challenge of increased hardware footprint and synthesis complexity when compared to recurrent models, where the key components can get reused to meet the resource constraints. To better understand the transformer's complexity, the next subsections derive the equations linking the internal components and explain the involved terminology.

3.1.1 Input embedding and Residual Connections

Although the model lacks any recurrency, the transformer includes two residual connections which have been widely adopted since their successful application in ResNet neural networks [53]. They offer improvements to training time and resulting accuracy [54], however, they require standardized data dimensionality to ensure the summation can be logically executed. In this project, this is obtained thanks to input embedding, which transforms the input $\hat{\mathbf{x}}^i \in \mathbb{R}^{L \times 16}$ into a shape $\mathbb{R}^{L \times d}$ that is used through the design, as seen in equation 3.1. The d size is referred to as the self-attention latent or embedded dimension.

$$\hat{\mathbf{x}}_{\text{emb}}^i = \text{embedding}(\hat{\mathbf{x}}^i) = w_{\text{embed}} \hat{\mathbf{x}}^i + b_{\text{embed}} \in \mathbb{R}^{L \times d} \quad (3.1)$$

This dimensionality change can be conveniently performed using a linear layer, and it has to be remembered that each such layer increased the model learning capacity thanks to the learnable weights and bias. The network's inner dimension d is treated as a hyperparameter as it influences the model's accuracy and performance, but it has to be noted that the other dimension prevalent in the network comes from the input's number of jet constituents L (which is set to 1 in case of the HLS representation), meaning that the model is also susceptible to a parameter which cannot be easily tuned.

3.1.2 Input Encoding

Along the embedding, an input encoding is concatenated and fed to the transformer layer. In natural language processing, the encoding is meant to allow the model to benefit from the sequential information of the words in a sentence. It can be obtained from a sinusoidal function using the position index or simply treated as another learnable parameter. The sequential relations are not present in the jet data, because all the jets originate from the same proton-proton collision, hence, the latter approach is used in this project. It is worth mentioning, that from empirical analysis, the learnable encodings have a significant impact on the final results as they represent a trained, hidden state concatenated to all inputs during evaluation, as shown in equation 3.2. Its impact is especially prevalent for the HLF data (where $L = 1$), where the hidden state matched input's dimension and effectively doubles it after concatenation.

$$\text{encoding}(\hat{\mathbf{x}}_{\text{emb}}^i) = w_{\text{encoding}} \in \mathbb{R}^{1 \times d} \implies \text{concat}(\hat{\mathbf{x}}_{\text{emb}}^i, w_{\text{encoding}}) \in \mathbb{R}^{(L+1) \times d} \quad (3.2)$$

Choosing a learned hidden state is also more efficient for inference in hardware, as the increased training cost associated with back-propagation of this parameter yields a constant set of values that are known during compile-time of the FPGA and can be implemented using a LUT.

3.1.3 Normalization and Parameter Extraction

As layer normalization does not track and gather running mean and variance statistics, this mechanism is implemented on top of the existing PyTorch implementation to facilitate extracting the aggregated statistics after training. These, along with all the learned weights and biases, are extracted and transformed into specific C++ formats supported in HLS using a custom tool developed for this purpose. This allows for directly initializing the FPGA's BRAMs and LUTs with the model parameters, which avoids the need for an interaction with a host machine.

Obtaining the statistics taken for the data before normalization layers can also be viewed as a hardware-aware optimization. This can be explained with the mathematical derivation presented in equation 3.3

$$y = \frac{x - E[x]}{\sqrt{\text{Var}[x] + \epsilon}} \cdot \gamma + \beta = x \cdot \left(\frac{\gamma}{\sqrt{\text{Var}[x] + \epsilon}} \right) + \left(\beta - \frac{\gamma \cdot E}{\sqrt{\text{Var}[x] + \epsilon}} \right) = w \cdot x + b \quad (3.3)$$

By treating the mean $E[x]$ and variance $\text{Var}[x]$ of input x as learned parameters, the square root and division operations can be fully omitted by fusing them into the existing γ and β parameters which simplifies the hardware required for the normalization layers. This is especially useful as

FPGAs lack dedicated hardware for these computationally expensive operations, which could lead to suboptimal designs being synthesized. Independently of the implementation in this work, a similar idea has been proposed and successfully used as an optimization in the past [55].

The algorithm behind the parameter extraction is rather simple, and the difficulty comes from the domain specific knowledge of handling PyTorch model parameters and generating the correct files for HLS. The break-down of the necessary steps can be seen in algorithm 3.1.

Algorithm 3.1 Mechanism behind model parameter extraction

```

1:  $state \leftarrow \text{load\_state}(\text{model})$ 
2:  $\text{sort}(state)$ 
3:  $curr\_weight \leftarrow \text{null}$ 
4: for  $param$  in  $state$  do
5:    $mean \leftarrow \text{find\_mean}(\text{model}, param)$ 
6:    $var \leftarrow \text{find\_var}(\text{model}, param)$ 
7:   if  $param$  is weight then
8:      $curr\_weight \leftarrow param$ 
9:      $new\_param \leftarrow \text{update\_weight}(param, var)$ 
10:  else
11:     $new\_param \leftarrow \text{update\_bias}(param, curr\_weight, mean, var)$ 
12:  end if
13:   $\text{save}(new\_param)$ 
14: end for

```

3.2 Ultra-Low Latency Architecture

The first of the proposed architectures targets the HLF representation dataset, where each sample is of $\mathbb{R}^{1 \times 16}$ dimensions, making it a better candidate for an N times faster inference than the constituent list representation with inputs with $\mathbb{R}^{N \times 16}$ dimensions. The simpler data means that the network can achieve satisfactory classification results with a lower learning capacity, hence allowing for various simplifications of the base architecture.

3.2.1 Optimization and Tuning

The hypothesis leading the optimization process is that the HLF dataset can be learned by a relatively lower complexity network. The strategy is to start with the accuracy and AUC value of the base architecture and keep applying changes to the network as long as they do lead to significant drops in classification results, with detailed evaluation discussed in chapter 5. Doing so is inherently beneficial for the later FPGA mapping, as even without hardware-aware changes, a simpler model is likely to require fewer resources and achieve lower latency.

The architecture features that have the highest influence on the overall performance are the size of the latent dimension and number of transformer layer. In the case of the original transformer model [31], 12 transformer layers are used, split equally into a decoder and encoder parts of the networks. This, along with latent dimension size of 512 shows that jet tagging requires significantly less complexity compared to natural language processing, as the recent ConstituentNet [8] is based on only 3 transformer layers and embedded dimension of 64. However, as the latter network targets a more complex dataset, the first optimization performed in this project yielded a reduction to only a single transformer layer with latent size of 16. The network complexity is directly proportional to the number of layers, while the self-attention complexities shown in section 2.2.4 can be simplified given the dimensions of *queries*, *keys*, and *values* are shared and equal to $\in \mathbb{R}^{N \times d}$ in this work, as shown in equation 3.4 and 3.5 for time and space respectively. It is worth mentioning that the shared dimension d is purposefully represented this way as it is equal to the model embedded dimension d .

$$\mathcal{O}(h \cdot (N_Q N_K (d_Q + d_V) + d_Q^2 (N_Q + N_K) + d_V^2 N_K + N_Q d_V d_{out}))) \equiv \mathcal{O}(h N d (N + d + d_{out})) \quad (3.4)$$

$$\mathcal{O}(h \cdot (N_Q(N_K + d_V) + d_Q(N_Q + N_K) + d_V N_K + N_Q d_{out}))) \equiv \mathcal{O}(hN(N + d + d_{out})) \quad (3.5)$$

Another successful optimization was simplifying the SiLU to ReLU activations, as in hardware, the latter rely on a single comparator to determine whether to pass input to output directly or set the signal to zero. In contrary, SiLU requires computing the sigmoid function as well as performing a multiplication, which has a non-trivial cost on an FPGA.

Interestingly, the layer normalization was first changed to batch normalization to simplify the tracking and embedding the running data statistics. This only leads to a marginal accuracy decrease, however, it is interesting to point out that so does removing the normalization layers from the design. This can be attributed to the "smoother" distribution of elements of the normalized HLF dataset, which tend to average around zero instead of taking values from a wide range as it is the case for the constituent list representation, shown in figure TODO. Hence, the used dataset does not benefit from additional normalization, vastly simplifying the architecture.

reference
figure from
physics
back-
ground
when it is
done

3.2.2 Summary

The optimizations described in the previous subsection correspond to a noticeably simpler architecture, which can be seen in figure 3.2. It is also worth mentioning that the simplifications targeted the components of the transformer block other than the multi-head self-attention block. This is a result of the used mechanism's key role in the learning process as well as the fact that it is difficult to modify it as it uses very basic components, and changing them would impact the underlying, commonly-approved function.

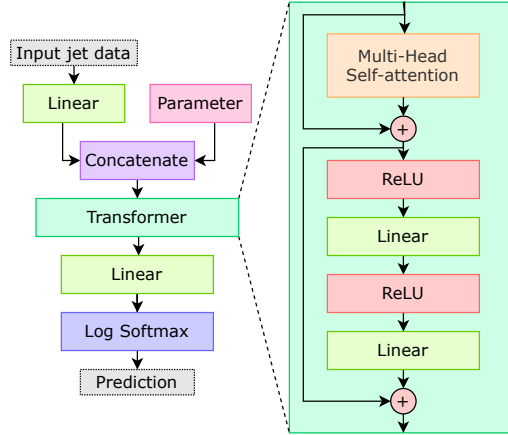


Figure 3.2: Diagram with an overview of the ultra-low latency architecture.

3.3 Accuracy-Focused Architecture

The aim of the second architecture is to exploit a bigger design along with the more complex constituent list dataset to achieve a higher accuracy and AUC value than the ultra-low latency model while still meeting the latency constraints of L1T at LHC. In addition, this design undergoes a design space exploration to compare and reason about a number of existing and novel techniques.

3.3.1 Optimization and Tuning

Aside from the increase in computational complexity, the biggest difference caused by changing the datasets is the vastly different range and distribution of samples, showcased in figure TODO. While the network could perform normalization as its first layer, this approach was tested to significantly degrade the results quality. This is likely due to the fact that each sample's feature follows a

distinct distribution and lays within a different range, so normalizing causes a loss in the carried information.

Normalization and data distribution also need to be discussed for a different reason in the context of this architecture. The layer normalization layers present in the base design are not completely omitted this time, and instead changed to batch normalization. Although the layer normalization characteristic of not tracking running statistics and relying on per-layer normalization during evaluation can yield improvement to classification results, it is not a desired behavior for FPGA mapping due to the lack of hardware support for the square root and division operations. This could be resolved in an analogous way to batch norm by simply manually tracking the statistics, but this was experimentally proven to substantially affect the accuracy. What is more, the less clustered and more diverse distributions of constituent list dataset features lead to numerical instability in softmax calculations of the exponential function. Even though this was a marginal issue for the PyTorch implementation, it was quickly observed to drastically change the expected outcome of the early HLS prototype implementations. A solution to this issue lays in using additional normalization layers just before the self-attention softmax blocks, which not only solved this problem, but also very slightly increased the learning capacity of the network.

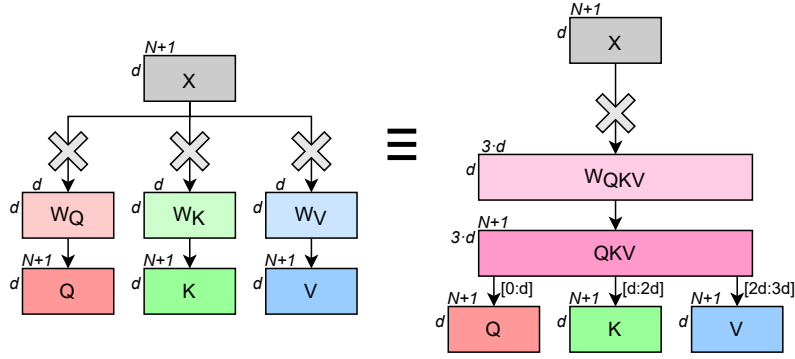


Figure 3.3: Simple illustration for equivalency between matrix multiplication using merged and unmerged weights on the example of *queries*, *keys*, and *values* used in self-attention.

It is a common for GPU training to merge together adequately shaped fully-connected layers, as seen in figure 3.3. This allows the GPU to more optimally use its resources for the computation, potentially decreasing the training time without affecting the results. Nonetheless, this approach was dropped for this architecture to avoid generating one bigger weight matrix in favor of several smaller ones. Contrary to GPUs, the synthesis process for FPGAs benefits from smaller pre-computed matrices as it is easier for the algorithm to map them to certain BRAMs or LUTs that are close to the components that read from or write to them. With the significantly smaller dimension sizes for the ultra-low latency model, this aspect generated negligible difference, while in this architecture the merged weight matrix size would exceed a reasonable size and lead to an extremely long synthesis process and likely suboptimal hardware mapping.

Lastly, some optimizations are inherited from the ultra-low latency architecture. The SiLU are replaced by ReLU activations, while embedding and encoding are simply implemented by a fully-connected layer and a learnable parameter. When it comes to the transformer layer count and latent dimension size, they follow the configuration of ConstituentNet of 3 and 64, however, a closer evaluation and discussion about these are carried in chapter 5.

3.3.2 Summary

In this case, the resulting architecture is much more similar to the base one in terms of its overall structure, however it differs from it with the crucial changes to normalization layers, activation function and modifications to the multi-head self-attention block. The overview of this design can be seen in figure 3.4.

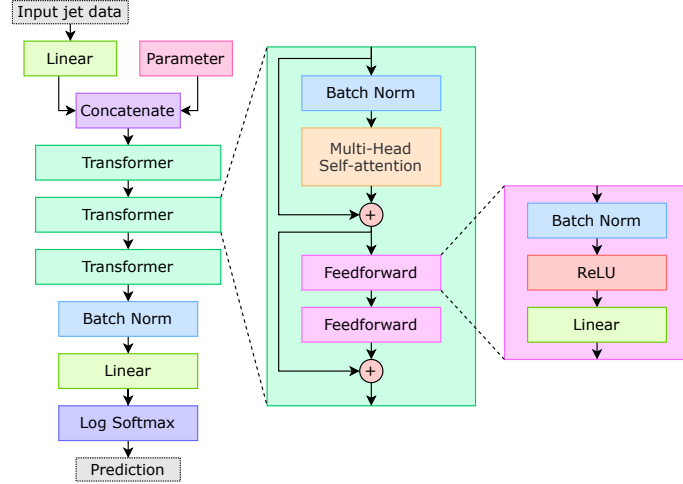


Figure 3.4: Diagram with an overview of the accuracy-focused architecture; the inclusion of batch normalization before softmax within the multi-head self-attention block is not showcased for the sake of clarity.

3.4 Quantization-Aware Training

The use of the term *quantization* in the topic of machine learning and reconfigurable hardware might appear as a misnomer if one follows the definition used in signal processing, where it refers to constraining continuous values to a set of discrete ones. However, this definition can be extended for converting values from a very large set of discrete values to a significantly smaller ones. This is the case when half (16 bits), single (32 bits) or double (64 bits) precision floating-point numbers are expressed using a fixed-point representation. This leads to an inherent quantization noise because many distinct values in floating-point representation map to the same number when stored as fixed-point, which can be seen in figure 3.5 [56]. Although this work exploits fixed-point arithmetic for its efficient hardware representation, it has to be noted that quantizing is also used for time and space efficient GPU inference as it can vastly lower the total bits used across model's weights, hence reducing the required memory and computational complexity.

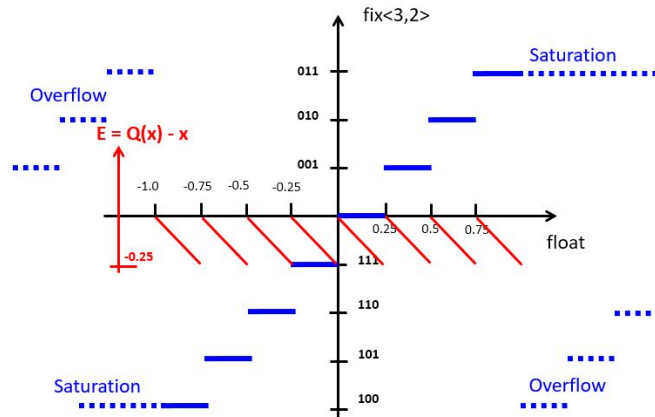


Figure 3.5: Quantization noise from 32-bit floating-point to 3-bit (2-bit integer, 1-bit fractional) fixed-point representation; widths selected to highlight the behavior. Two methods of handling values outside representable range are showed: classic overflow as well as saturation to minimum or maximum value.

3.4.1 Floating-point and Fixed-point Representations

To bring more perspective to this topic, let's start by observing the 16-bit wide floating-point notation with an example seen in equation 3.6. *SIGN*, *EXP* and *FRAC* stand for the sign, exponent and fractional³ fields (sequences of 1's and 0's) of a number.

$$\{\text{SIGN} \mid \text{EXP} \mid \text{FRAC}\} = \begin{cases} (-1)^{\text{SIGN}} \cdot 2^{-14} \cdot (0.\text{FRAC}_2) & \text{EXP} = 00_000_2 \\ (-1)^{\text{SIGN}} \cdot 2^{\text{EXP}-15} \cdot (1.\text{FRAC}_2) & \text{EXP} \in [00_000_2, 11_110_2] \\ \pm\infty & \text{EXP} = 11_111_2 \end{cases} \quad (3.6)$$

The widths of the fields for standard floating-point width can be seen in table 3.1, however, the particular details in the example do not need to be fully understood to notice the required complexity of the hardware design to perform operations using floating-point values. It has to first decode each of the fields, check for special values, normalize inputs to a common base, perform the computation, and lastly normalize again and decode the fields. As it can be expected, this comes with a significant overhead compared to integer numbers.

Table 3.1: Comparison of standard precision types for floating-point values. Note the difference in half precisions, where *Brain* offers same the exponent width as single precision, and *binary* priorities fractional precision.

Precision	Total width	Field width		
		sign	exponent	fractional
Double	64	1	11	52
Single	32	1	8	23
Half (<i>Brain</i>)	16	1	8	7
Half (<i>binary</i>)	16	1	5	10

While dedicated circuits exists in CPUs and GPUs to accelerate floating-point calculations, FPGAs tend to use fixed-point representation to achieve superior latency with relatively simple hardware resources. It comprises a sequence of P binary digits representing the integer part, then a virtual decimal point, followed by R binary digits representing the fractional part, according to equation 3.7.

$$\{I_{P-1}|\dots|I_0|F_{R-1}|\dots|F_0\} = I_{P-1} \cdot 2^{P-1} + \dots + I_0 \cdot 2^0 + F_{R-1} \cdot 2^{-1} + \dots + F_0 \cdot 2^{-R} \quad (3.7)$$

The virtual decimal point is not explicitly present in the number, instead the lengths of the sequences are stored which allows for correct handling. There are also no widely adopted standards for the internal widths of fixed-point numbers, as they are often implementation-specific. Fixed-point hardware is only responsible for correctly aligning the inputs in case they vary in width, which is an information know at compile-time, as opposed to run-time for floating-point numbers, which makes it very suitable for FPGAs.

3.4.2 Existing Implementations

Quantization-aware training refers to training a model using the fixed-point representation of numbers to allow the model to reduce the impact of the quantization noise by learning the underlying data characteristics with less precision. However, approaches differ in the degree to which the quantization is used as it can be applied to training data, weights and biases, intermediate results or any combination of these. It also has to be noted that existing implementations have been designed with different objectives, although to the best of our knowledge, the most widely used ones are still in their very experimental phases at the time of writing this report. The approaches considered for this work are listed below:

- **PyTorch Eager Mode Quantization** [57] - currently in a *beta* version. PyTorch offers two development modes - Eager, for research purposes and easier experimentation, as well as

³Also referred to as *significand precision*

Script, for production use cases [58]. This quantization scheme is an extension to the former mode and requires manual handling of the placement of quantizing and dequantizing layers in a network as well as operator fusion and functional blocks. It currently offers no support for the vanilla recurrent block, LSTM, GRU or self-attention.

- **PyTorch FX Graph Mode Quantization** [59] - similarly to the previous point, it is in a *prototype* version, and it extends the graph-base and speed-focused Script mode. This leads to an easier implementation as the quantization feature support process happens automatically. It does not work for convolutional, self-attention or embedding layers yet.
- **Brevitas** [60] - a research project from a team at Xilinx, a major reconfigurable hardware vendor. It is early in its development cycle, but it promises to support most network architectures, provided that all existing layers are replaced with their quantized variants, and quantizing and dequantizing is also handled manually, similarly to PyTorch Eager Mode Quantization. Its caveat is that while it provides a granular, per-variable level of bit-width customization, the offered fixed-point widths cannot be easily changed in size.
- **QPyTorch** [61] - a Cornell University research project, which offers a fixed-point simulation framework. Existing networks can be easily adapted to it, only requiring manual quantization layers usage throughout the design. These can be of arbitrary width, which supports a rich design space exploration, however, the representation of widths, biases, accumulators and gradients needs to follow the same quantization scheme for all layers, which is dictated by the training optimizer configuration.

It is clear that each described solution comes with its own advantages and disadvantages. Even though the first two approaches come officially from PyTorch, they lack support for the core transformer layer - the self-attention. Hence, this report proposes the latter two for research on quantized transformer architectures and their hardware implementations.

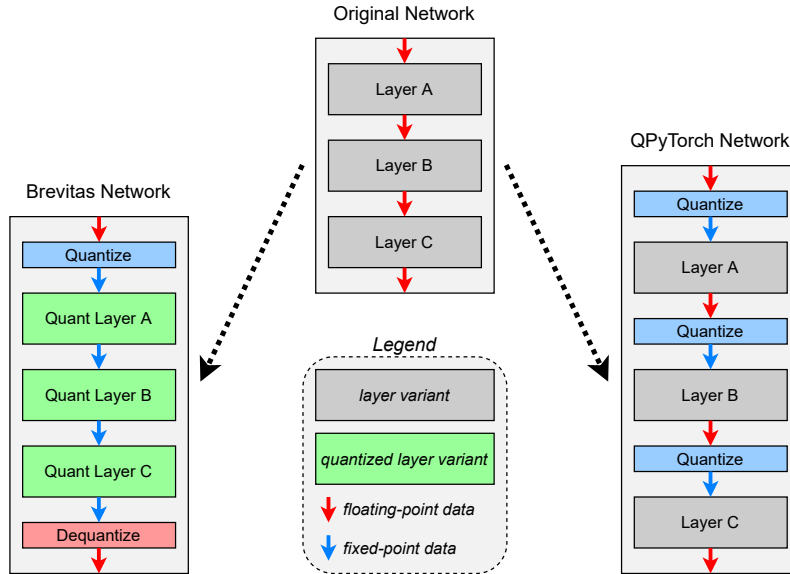


Figure 3.6: Comparison of a simple neural network prepared for quantization-aware training using Brevitas and QPyTorch frameworks.

Figure 3.6 showcases the designs resulting from quantizing a simple neural network using Brevitas and QPyTorch. The former approach relies on library-provided quantized variants of commonly found neural layers, and each can have its input's, output's and internal parameters' bit-width configured separately to one of the pre-defined types. This technique is more truthful to how FPGAs operate, but at the same time, more demanding in terms of implementation complexity than QPyTorch's one, which exploits the wide range of PyTorch's native layers and only requires custom quantizing blocks in between them. The latter is also slightly less computationally intensive, however both methods results in significantly longer training time than the floating-point

representation due to the lack of support for fine-tuned, fixed-point GPU implementation of the network layers. It is also worth mentioning, that Brevitas offers a convenient way of exporting the quantized models to Xilinx hardware representations as well as external tool chains. Lastly, quantization is also proposed and discussed as a post-training mechanism in section 4.4, with both approaches then evaluated and compared in chapter 5.

Chapter 4

Hardware-Aware Analysis

In this chapter, the hardware-aware optimizations used in the FPGA implementations are first presented and then their application in the proposed architectures is explained. Other elements of the hardware design and configuration process are then described which results in a comprehensive picture of the FPGA-mapped architectures. Afterwards, two analytical models are considered - one for the latency and the other for the resource utilization. The custom post-training quantization tool is discussed along with its suitability for this project. Lastly, the technical contributions to `hls4ml` library are listed and explained.

4.1 Hardware-Aware Optimizations

4.1.1 Tensor Multiplication and Scaling

Each self-attention head performs two tensor multiplications (referred to as *matmul* blocks in figure 2.4), which are normally expressed using Einstein Summation notation [62], which is supported by mathematical and machine learning libraries like NumPy or PyTorch. However, not present by default in HLS, it requires careful design of the calculation loops in order to not cripple the performance by unnecessary computations and pseudo-random data accesses. As part of this research, an efficient and fully-customizable HLS block has been designed, that uses a very similar interface to the Python equivalent.

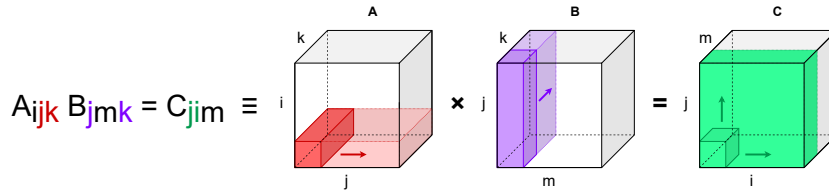


Figure 4.1: Visualization of a tensor operation expressed in Einstein Summation notation.

Figure 4.1 shows a visualization for an example notation to give a better understanding of the necessary flexibility of a formula. The translation between notations using the custom tool is showcased in listing 4.1. While the PyTorch implementation can often use 4-dimensional tensors, the first dimension refers to the batch, which is not present in the hardware implementation that processes input samples one-by-one, hence both the figure and code listing show 3-dimensional cases. It is also worth pointing out, that tensor multiplication is an inherently computationally expensive operation due to the quadruple-nested loop structure. For this reason, the proposed design leaves the configuration of pipelining as a parameter that offers a trade-off between time and design complexity. The *design* complexity refers to the difficulty involved in HLS synthesis as well as hardware resource utilization. In other words, the hardware block can be instantiated to run serially, where little resources are needed as they get re-used, or alternatively, in parallel,

where significantly more components are used to decrease latency, and in case the design is also pipelined, to also increase throughput.

```
#define EINSUM_PIPELINE(DEPTH) {\
    if (io_type == io_parallel) {\
        if (pipeline_einsum_depth == DEPTH) {\
            #pragma HLS PIPELINE II=reuse_factor\
        } else {\
            #pragma HLS unroll \
        }
    }
}

type_t out[H][K][Q];
if (io_type == io_serial) {
    #pragma HLS ARRAY_PARTITION variable=out cyclic factor=cycle_factor
} else if (io_type == io_parallel) {
    #pragma HLS ARRAY_PARTITION variable=out complete
}

Einsum_H: for (unsigned hh = 0; hh < H; hh++) {
    EINSUM_PIPELINE(3);
    Einsum_K: for (unsigned kk = 0; kk < K; kk++) {
        EINSUM_PIPELINE(2);
        Einsum_Q: for (unsigned qq = 0; qq < Q; qq++) {
            EINSUM_PIPELINE(1);
            Einsum_C: for (unsigned cc = 0; cc < C; cc++) {
                EINSUM_PIPELINE(0);
                if (cc == 0) out[hh][kk][qq] = 0;
                out[hh][kk][qq] += A[qq][hh][cc] * B[kk][hh][cc];
            }
        }
    }
}
```

List of Listings 4.1: From PyTorch `out = torch.einsum("qhc,khc->hqk", [A, B])` to HLS C++ code.

$$\text{Design : } \mathcal{O}(n) \quad \text{Time : } \mathcal{O}(HKQC \cdot n) \quad (4.1)$$

Let's consider the two extreme cases for the design - no loop unrolling and complete unrolling, where the latter is required for the block to be fully pipelined, and assume that multiply-accumulate and addition both have n time and space complexity. In the first case, a single multiply-accumulate operation happens at once, hence a final result is only available after all the loops have been fully iterated, with complexities shown in 4.1.

$$\text{Design : } \mathcal{O}(HKQC \cdot n) \quad \text{Time : } \mathcal{O}(n \cdot \log(HKQC)) \quad (4.2)$$

In the second one, all loop operations can execute at the same time, although the intermediate results need to be summed accordingly using an adder tree (seen in figure 4.2) which has a logarithmic time and linear space complexity, before saving the output tensor, which leads to complexities seen in 4.2. The time complexity may appear high, but it has to be remembered that unrolling allows for the pipelining of this design, which cannot decrease latency, but can vastly increase throughput, as each addition can happen in a single cycle.

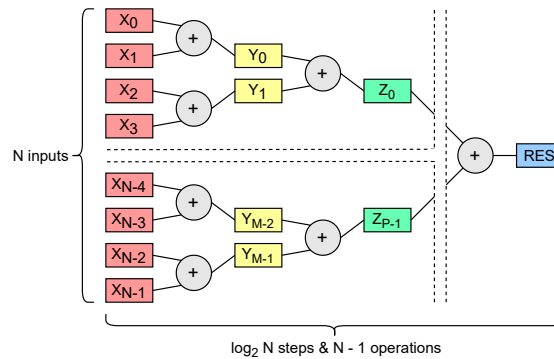


Figure 4.2: Illustration of an adder tree with N inputs.

Another simple optimization used alongside the tensor multiplication blocks was the change in size scaling from using division to performing an arithmetic right shift (ASR), which requires precomputing the logarithm of the size, seen in equation 4.3, vastly simplifying the otherwise computationally expensive hardware required at run-time.

$$\frac{x}{\sqrt{\text{size}}} \equiv \text{ASR}(x, \log_2 \sqrt{\text{size}}) \equiv \text{ASR}(x, \frac{1}{2} \log_2 \text{size}) \quad (4.3)$$

4.1.2 Softmax and Log Softmax Activations

Despite an already existing `hls4ml` implementation of the softmax activation function, obtaining the logarithm of its result is not as simple as it may seem. This is because the numerical stability and computational efficiency of this operation is non-trivial and varies depending on the programming language and target platform, as it is a topic of ongoing research [63].

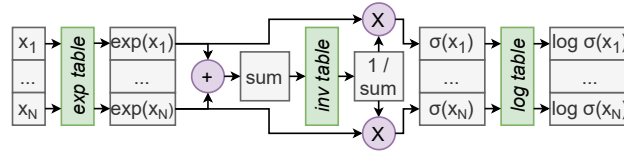


Figure 4.3: Direct hardware implementations of log softmax.

The naive implementation comes straight from the definition of taking a logarithm of softmax, seen in equation 4.4, and the required hardware operations are shown in figure 4.3.

$$\sigma(x_i) = e^{x_i} / \sum_{j=1}^N e^{x_j} \quad (4.4)$$

This report proposes a different way of mapping this operation to hardware to improve stability while shortening the critical path and using less resources. It is based on the derivation shown in equation 4.5.

$$\log(\sigma(x_i)) = \log(e^{x_i} / \sum_{j=1}^N e^{x_j}) = \log(e^{x_i}) - \log(\sum_{j=1}^N e^{x_j}) = x_i - \log(\sum_{j=1}^N e^{x_j}) \quad (4.5)$$

The resulting hardware operations are depicted in figure 4.4. It is important to note, that operations like exponentiation, division or taking a logarithm usually rely on precomputing a wide range of values and mapping them in BRAMs or LUTs to allow for lookup on run-time. Hence, the optimized design requires one less of such lookups while also replacing multiplication by a subtraction, which can be simpler to express in hardware.

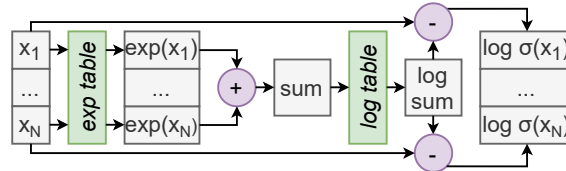


Figure 4.4: Optimized hardware implementations of log softmax.

Although further simplifications, including approximating the summation by finding the maximum (see equation 4.6) or simply omitting the logarithm portion of the expression, were also explored, they noticeably lowered the final accuracy and were thus abandoned.

$$\log(\sigma(x_i)) = e^{x_i} - \log\left(\sum_{j=1}^N e^{x_j}\right) = e^{x_i} - \sum_{j=1}^N \log(e^{x_j}) = e^{x_i} - \sum_{j=1}^N x_j \approx e^{x_i} - \max(x) \quad (4.6)$$

4.2 Neural Networks Implementation

The FPGA implementations of the two proposed architectures capture the ideas and novel approaches explained in chapter 3 and the previous subsection. What is left to discuss are the particular methods used in HLS C++ that play a significant role in the models' performance and quality of the results.

As its name suggests, the ultra-low latency architecture is designed to offer cutting-edge inference speed. Thus, its hardware implementation is fully parallelized and pipelined, which means that operations are mapped to separate hardware resources split into pipeline stages wherever it is feasible. To achieve this, all loops are unrolled and every array⁴ is fully partitioned in order to be stored in registers instead of slower-to-access BRAMs. This not only avoids any loop overhead, but also allows the computations to happen at the same time regardless of the design size, provided enough hardware resources are available.

The second, accuracy-focused architecture relies on a more complex dataset, hence requiring a substantially bigger design. This limits the possibility of performing every calculation in parallel and introduces the need to reuse the most demanded components. To avoid resource bottlenecks, arrays can no longer be completely partitioned to reduce the number of used FFs and LUTs in favor of BRAMs. For the same reason, certain computationally-expensive loops cannot be unrolled to restrict immense DSP usage. Surprisingly, the complexity of the design's synthesis process has to also be taken into consideration, which is a limitation of the existing HLS technology. This challenge imposes new constraints on the model's configuration, which is discussed in details along its evaluation in chapter 5.

4.3 Analytical Latency and Resource Models

In the majority of cases, simulation and synthesis are the best ways of measuring model's accuracy, latency and hardware resources utilization. Although the results might vary compared to an actual FPGA implementation due to various nuances involved in converting the RTL code to equivalent bit stream, the overall design trends remain unchanged. Nonetheless, deriving an analytical model offers a much simpler and easier way of understanding the consequences of modifying a design. While predicting accuracy is not a feasible endeavor due to the complex nature of neural network models, latency estimation is possible for a given design configuration. Tracking hardware resources is not manageable in general as used FF and LUT numbers can reach millions, however, rough estimation of DSP slices can be discussed. In order for the analytical models to be easier to understand and verify, only the smaller, fully pipelined, ultra-low latency architecture is discussed in this section.

4.3.1 Latency Model

A visualization of the architecture annotated with the pipeline stages can be seen in figure 4.5. As a consequence of pipelining the model, all the operations can be completed in either one or two cycles. Certain operations like ReLU or scaling share their stage with subsequent components due to their simplicity.

The following components' stages last for two cycles, which suggests that their internal operations could not be completed within a cycle:

⁴Multidimensional arrays are used in HLS C++ to implement matrices and tensors in the most efficient ways.

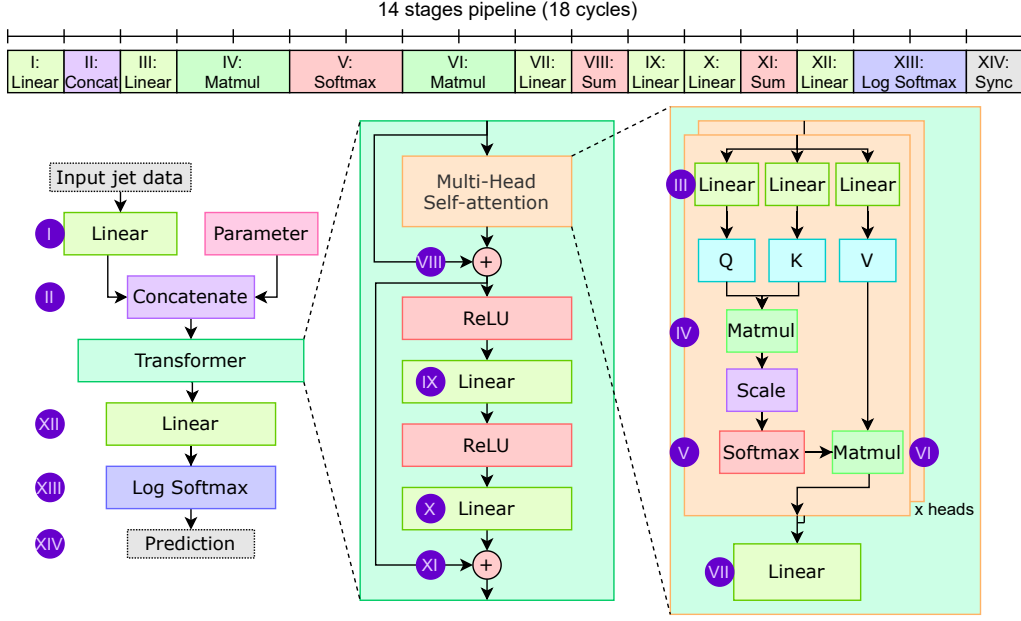


Figure 4.5: Ultra-low latency architecture with highlighted pipeline stages. Not annotated blocks share stage with their neighbors.

- **Tensor product** - a single multiply-accumulate operation requires only one cycle, which might seem counterintuitive at first. However, in this case, several multiplication results need to be added together for each output's value, which leads to additional complexity and latency.
- **Softmax activation** - it performs an exponential function lookup for each input in parallel given the full partitioning of the arrays, but then the values need to be summed, and more importantly, multiplied, which generates the additional cycle.
- **Log softmax activation** - compared to the previous operation, no multiplication is needed in this case, saving a cycle. However, after the initial lookups, another one needs to happen to find the logarithm value, which requires an additional cycle.

The design configuration affects several of the pipeline stages, and hence a significant portion of the latency. The number of transformer layers have a direct impact on stages III to XI, with each additional layer effectively duplicating them, increasing the latency by 12 cycles. Within the transformer block, each feed forward (linear and ReLU) layer contributes one cycle. It is not obvious at first, but the number of self-attention heads do not influence the latency as they are executed in parallel.

$$\text{latency (cycles)} = C_{top} + T \cdot (C_{SA} + C_T + F \cdot C_{FF}) \quad (4.7)$$

Overall, the latency can be expressed using a few variables and constants, leading to the dependency seen in equation 4.7. As for the notation, C_{top} , C_{SA} , C_T , and C_{FF} stand for the constant cycles needed in the top-level, self-attention, transformer, and feed-forward components, while T and F mean the number of transformers and feed-forward layers.

$$\text{latency (cycles)} = 6 + 1 \cdot (8 + 2 + 2 \cdot 1) = 18 \quad (4.8)$$

Given the design configuration of the ultra-low latency model, 18 cycles are expected as the overall latency value, as proven in equation 4.8. The number of transformer layers plays the most important role in this model, which can be seen from the plot in figure 4.6 which shows a 3-dimensional visualization of the latency dependency, with the model's constant values.

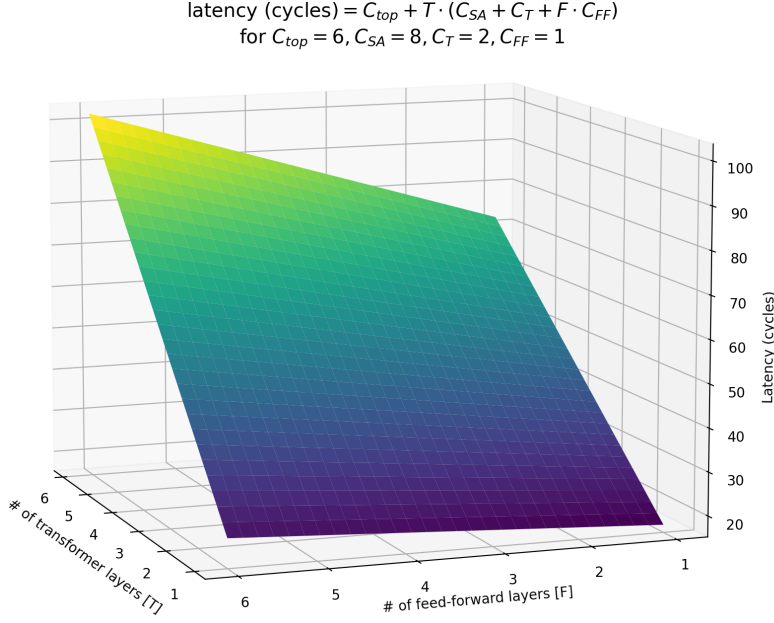


Figure 4.6: 3D plot of the latency model dependency on the number of transformer and feed-forward layers.

4.3.2 DSP Model

The raised abstraction level of C++ code compared to an RTL representation is the main benefit of HLS, but it poses a challenge for accurate modelling of hardware resources. Hence, the DSP model uses precise values wherever feasible, and big O notation for the more convoluted components.

$$\text{Linear: } \# \text{ of DSPs} = \mathcal{O}((\text{input size}) \cdot (\text{output size})) \quad (4.9)$$

Starting with the ones that are not expected to contribute any utilization, both ReLU and scaling can be implemented without DSPs, using solely comparators for the conditions and clever routing for shifting. While softmax activation requires one multiplication for each of its inputs which depends on *# of input jets*⁵ called L , the optimized log softmax uses subtraction instead and can be omitted in this model. The most difficult to model are the more computationally expensive components - linear layers and tensor multiplication blocks. The former performs a number of multiplications proportional to its input and output sizes, directly relating to DSP count, as well as an accumulation of the products which is not guaranteed to be mapped to separate units, possibly using existing ones instead. The resulting complexity can be seen in equation 4.9.

$$\text{Tensor product: } \# \text{ of DSPs} = \mathcal{O}(L^2 \cdot H) \quad (4.10)$$

Similar argument can be made about the tensor multiplication, where it is also clear that there is a direct relation to the size of the involved tensor dimensions, but it is not easy to conclude how many DSP slices each multiply-accumulate need. Given that both of the instances of this block operate on the same dimensions, equation 4.10 uses a more precise notation, where H stands for the number of self-attention heads.

In order to express the overall DSP model, the input and output sized of each linear layer have to be considered. The following list refers to them using the pipeline stages introduced in figure 4.5 for unambiguity:

- **Stage I:** from jet feature number (16) to self-attention latent dimension d ,

⁵The term *# of input jets* refers to the dataset's jet dimensionality - 1 for HLF and 30, 50, 100, or 150 for the constituent list representation.

- **Stage III:** regardless of being fused into 1 or expressed as 3 separate layers, it takes the latent dimension d and yields $3 \cdot d$,
- **Stage VII:** both input and output are of size d ,
- **Stage IX:** from d to $2 \cdot d$,
- **Stage X:** from $2 \cdot d$ to d ,
- **Stage XIII:** from d to the number of classes (5).

All together, the above dimensions can be summed to $16d + 3d^2 + d^2 + 2d^2 + 2d^2 + 5d = 21d + 8d^2$. To produce a clearer model, constant coefficients C_{linear} and C_{tensor} are introduced in the computational complexities to express them directly. Hence, combining the above information about softmax, linear and tensor product blocks, the following DSP model shown in equation 4.11 is obtained. As expected, the result is significantly more complex than the latency model and cannot be easily plotted.

$$\# \text{ of DSPs} = L + C_{\text{linear}} \cdot (21d + 8d^2) + C_{\text{tensor}} \cdot 2 \cdot L^2 \cdot H \quad (4.11)$$

4.4 Post-training Quantization

This section returns to the topic of quantization, but as opposed to section 3.4, it explores the quantization of already trained models. This domain is not researched as much as quantization-aware training due to the lack of ability for a model to *compensate* for the quantization noise during training, hence leading to potentially inferior results. However, recent advancements in this field [64] leverage the synergy between post-training quantization and the target hardware platform to produce results with improved latency or energy consumption. The inherent noise issues are offset by a careful per-variable bit-width analysis, driven by a reinforcement learning algorithm. The choice of the algorithm has a deep-rooted issue for more computationally demanding models that also require a search in a wider range of bit-widths⁶.

4.4.1 Motivation

This report proposes a novel post-training quantization algorithm that can be applied to state-of-the-art transformer neural networks over a wide precision range. Early tests in the HLS environment revealed that a single C simulation for an input with only 100 samples can take around 10 minutes, which dictates a need for a significantly simpler, hence faster, algorithm than Bayesian optimization or reinforcement learning to allow for an exploration that runs in a reasonable amount of time.

The motivation of the algorithm comes from a hypothesis which states that the neighboring layers in a neural network have a relatively high correlation in their optimal bit-widths. Under this assumption, each layer's input, output, weight, bias and accumulator can be safely explored one-by-one, in the order of appearance in the model. *Safely* refers here to a low likelihood of arriving at a local accuracy extremum that is substantially worse than the global one, that could only theoretically be found using a more sophisticated approach. During this *walk* through the design space, several non-trivial constraints about the widths have to be ensured, which are the topic of the next subsection.

4.4.2 Constraints

The constraints of a network variable could theoretically be set arbitrarily to convey a high-level requirement of an experienced designer with a knowledge about typical widths used for a component in a given network type. However, the proposed method automates this process by extracting the underlying lookup table characteristics to accommodate users without domain-specific expertise. These characteristics are part of the network configuration that ensures that any precomputed (for increased latency) function is stored with adequate precision that avoids introducing unnecessary errors. To give a more concrete example to this abstract definition, one can consider the range of

⁶The mentioned method only explores convolutional neural networks in [1, 8] bit-width range.

values yielded from the exponential function. Not only is there a set width for the results, but even relatively small values map to numbers that require several bits of integer precision, so careless reduction to either of the width parts can quickly degrade any learning capacity of the model.

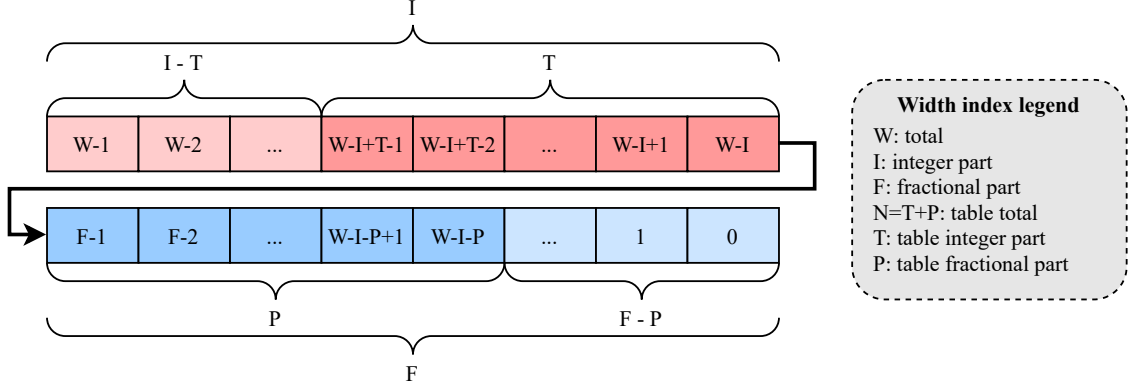


Figure 4.7: Visualization of a fixed-point number with its bit-widths as well as constraints imposed by a lookup table. For convenience, red and blue distinguish integer and fractional parts, while the darker hue shows the table-related parts.

Figure 4.7 visualizes a fixed-point number with a detailed analysis of its structure in terms of lookup table constraints, which were added as an optional feature, and are described in section 4.5.3. At first, it could be assumed that the imposed widths should simply be adopted by the variables used by the corresponding table values. However, it is possible for such variables to also have connections to other paths in a network, which can require more precision than the table, hence justifying the existence of the constraints ranges presented in equations 4.12, 4.13, and 4.14, with the notation coming from the corresponding figure.

$$W \geq N > T \geq 1 \quad (4.12)$$

$$W \geq I \geq T \geq 1 \quad (4.13)$$

$$W \geq F \geq P \geq 1 \quad (4.14)$$

4.4.3 Steps

A high-level overview of the steps used in the proposed method is depicted in algorithm 4.2. The only required parameters are the negative and positive accuracy tolerance, which are used to guide the search process. Increasing the negative tolerance relaxes the accuracy constraints ensured after each algorithm step in order to significantly reduce the overall bit-widths. On the other hand, a complete opposite objective can also be achieved as reducing the positive tolerance prioritizes quality of results over narrower bit-widths. Overall, these parameters can take an arbitrary combination of non-negative values that best expresses the exploration goals.

The algorithm starts by scanning an HLS file that includes all the fixed-point types used in a model, convert them into objects and organizes them into a First In, First Out (FIFO) queue, according to their original order. One by one, objects are popped from the queue and undergo the search process. Firstly, the bit-width that was found in the file is saved, the object inherits from the previous optimal bit-width (based on the correlation hypothesis), and the model is evaluated. Depending on whether the results stay within the accuracy tolerance, the object starts with either its original or the inherited configuration. Afterwards, the integer and fractional parts go through one or two loops - first, at each iteration the width gets incremented as long as it matches the target range. After the first loop, the second one, analogous but checking designs with fewer bits, is only entered if no significant improvements were found when increasing widths. This avoids evaluating same configurations unnecessarily, which goes in line with the overall aim to offer a relatively quickly-converging solution to the very multidimensional optimization problem. Once

all width objects reach their optimal state, the found configuration is saved and can be used in a further design space exploration.

It is important to point out that the aforementioned PyTorch Eager Mode and FX Graph Mode quantization schemes offer post-training quantization, but there are unsuitable for this work due to their lack of flexibility and support for the essential neural network layers. Instead, the HLS source directory is modified and tested in a C Simulation to measure accuracy. While the synthesis process could supplement the algorithm with an additional feedback signal about the precise hardware resource utilization, it takes much more than the simulation, and its results can be estimated to vary linearly with the total number of bit-widths. The only situation when such assumption does not hold is when a hardware component input width limit is reached because then a 1-bit increase requires an additional instance of that component to be instantiated. Ways of avoiding this issue are covered along with the evaluation of the results in section 5.3.2.

4.5 hls4ml Contributions

In this section, the technical contributions to the `hls4ml` library are descriptively presented, highlighting the areas that this work expands upon. Developed components are shown in figure 4.8, along with a number of existing components that were expanded upon or are used to draw comparison with. Each group is discussed in one of the following subsections, in an order of increasing design complexity.

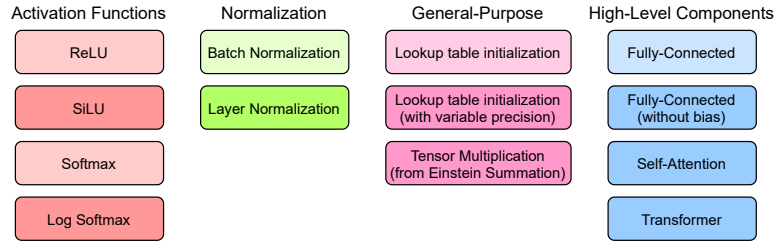


Figure 4.8: Overview of the created implementations (dark-colored) and some existing components with similar functionality (light-colored).

4.5.1 Activation Functions

The library offers several common non-linear activation functions, including the basic one - ReLU. An implementation was created for the SiLU, which extends the existing ReLU block by adding multiplication with the sigmoid function that can happen serially or in parallel depending on the specific configuration. In addition, the log softmax activation had implementations developed using both the direct and the optimized method described earlier in this chapter in section 4.1.2 to be able to evaluate the success of the optimization process.

4.5.2 Normalization Layers

Layer normalization and batch normalization perform the same operation across different dimensions, which on its own would not require a separate implementation. However, they also differ in terms of statistic collection, as only the batch version collects them during training to use during inference while layer normalization always computes them directly. While this is a suboptimal approach for reconfigurable hardware, a simulation-only version of layer normalization was implemented to allow for analyzing the effect of switching between these two methods to better understand their trade-off.

4.5.3 General-Purpose Blocks

The method used in `hls4ml` for mapping the precomputed function values to lookup table indexes prioritizes the integer range of a fixed-point number over its fractional part. In other words, it

Algorithm 4.2 Algorithm for performing post-training quantization search

```
function POSTTRAININGQUANTIZATION(neg_accuracy_tolerance, pos_accuracy_tolerance)
    previous_width  $\leftarrow$  null
    max_decrement  $\leftarrow$  neg_accuracy_tolerance  $\cdot$  2       $\triangleright$  Maximum decrement per parameter
    optimal_accuracy  $\leftarrow$  find_accuracy()
    params  $\leftarrow$  scan_file(defines_file)                   $\triangleright$  FIFO with scanned parameter objects
    while params not empty do
        current  $\leftarrow$  params.pop()
        if previous_width exists then                       $\triangleright$  Try using width from previous parameter
            original_width  $\leftarrow$  params.width
            update(params, previous_width)
            if find_accuracy()  $<$  optimal_accuracy - max_decrement then
                update(params, original_width)
            else
                optimal_accuracy  $\leftarrow$  find_accuracy()
            end if
        end if
        for part in {int, frac} do
            try_increase  $\leftarrow$  True
            pos_improvement_found  $\leftarrow$  False
            while try_increase do                           $\triangleright$  Increment to check for high accuracy gain
                param.increment(part)
                if find_accuracy()  $>$  optimal_accuracy + pos_accuracy_tolerance then
                    optimal_accuracy  $\leftarrow$  find_accuracy()
                    pos_improvement_found  $\leftarrow$  True
                else
                    try_increase  $\leftarrow$  False
                    param.decrement(part)
                end if
            end while
            if not pos_improvement_found then                 $\triangleright$  Decrement if no good increment
                try_decrease  $\leftarrow$  True
                acc_before_decrease  $\leftarrow$  optimal_accuracy
                while try_increase do
                    param.decrement(part)
                    if acc_before_decrease - find_accuracy()  $>$  max_decrement then
                        try_decrease  $\leftarrow$  False
                        param.increment(part)
                    else if find_accuracy()  $>$  optimal_accuracy - neg_accuracy_tolerance then
                        optimal_accuracy  $\leftarrow$  find_accuracy()
                    else
                        try_decrease  $\leftarrow$  False
                        param.increment(part)
                    end if
                end while
            end if
        end for
    end while
    return params
end function
```

simply uses the N most significant bits of a value for calculating the table index, where N is the table word width. Although this approach works in the majority of cases, it limits the ease of exploration as the type used in a lookup table might be affected by another variable’s width, hence enforcing an unnecessarily wider integer range and sacrificing decimal precision. A good example is the natural logarithm, which output range can be represented using fewer integer bits than its input domain, which makes it the perfect candidate for using the custom lookup table initialization. Aside from that, the tensor multiplication covered in details in section 4.1.1 was implemented in a way to match the style of existing tensor and array modifying components. This makes it very convenient for accelerating more complex designs that rely on Einstein Summation in their PyTorch implementation, and in the future, it could allow `hls4ml` to automatically translate that concept as it is currently not supported.

4.5.4 High-Level Components

It comes as no surprise that `hls4ml` offers a customizable and efficient fully-connected layer implementation. However, the way of creating an instance without the bias calculation is to provide an array filled with zeros. This is believed to be a shortcoming of the current version, as having a compile-time parameter regarding the bias can make the synthesis process faster given that the HLS can skip the optimization process for that aspect. With bigger and more complex models, this simple change could lead to noticeable reduction in processing time. What is more, it also yielded a marginally smaller design in the case of this work, suggesting that the HLS hardware translation might not be behaving optimally, but that depends on a number of factors and was not evaluated further.

Although the deeply covered self-attention and transformer layers could be viewed as specific to the topic of this report, it has to be remembered that they are the state-of-the-art when it comes to a plethora of applications. At the same time, they are incomparably more complex than any other aforementioned components, not only needing them for correct functionality, but also requiring considerable amount of development time to properly integrate and carefully optimize. Hence, they are considered the most important `hls4ml` contributions of this work.

Chapter 5

Evaluation

This chapter starts by evaluating the proposed neural network architectures on CPUs and GPUs to find a baseline inference latency, accuracy and AUC values. This information is then compared with the results obtained from simulating and synthesizing the models on reconfigurable hardware. The outcome of the design space exploration includes hardware resource utilization metrics as well as discussion about the Pareto front, applicability in high energy physics environments, and verification of the analytical models formulated in the previous chapter. Lastly, both the quantization-aware training and post-training quantization are evaluated quantitatively in terms of the trade-off between quality of results and bit-width reduction as well as qualitatively for their ease of adaptation to existing designs.

As efficient hardware mapping is the motivation behind the proposed models, out of the constituent list representation datasets, only the one containing 30 most energetic jets is analyzed in this chapter due to the inherent complexity of transformer neural networks and the challenges involved in mapping bigger designs to FPGAs.

5.1 Proposed Models' Performance

The first objective of the project is to design neural network architectures capable of very fast and accurate classification. Firstly, the classification accuracies and ROC curves are analyzed. Then, the choice of a dataset for both models is discussed, followed by inference latency results and comparison to existing solutions.

5.1.1 Classification Accuracy

Table 5.1 shows the evaluation results of the base and the two proposed architectures on the HLF and the constituent list datasets. Compared to the base one, the ultra-low latency model requires roughly 156 times fewer parameters and 2,875 times fewer FLOPS while achieving comparable classification results, suggesting that the base model has significantly too much learning capacity for that dataset. The second model also offers a notable parameters and FLOPS reduction of around 4 times. This matches the quadratic complexity of the self-attention layer given that the latent dimension is reduced from 128 to 64 and proves the importance of this factor. Along this, there is a slight improvement in accuracy and AUC thanks to a more careful selection of the design parameters.

It is interesting to observe which jet categories are harder to differentiate, which is summarized in the confusion matrix shown in figure 5.1. Both models tend to confuse the lighter particles, gluons with light quarks, although the ultra-low latency model is relatively better at classifying them. The other model achieves better results for the heavier particles - bosons and top quark.

5.1.2 Receiver Operating Characteristic Curves

Analyzing the ROC curves offers a more complex analysis of the trade-offs in classification results between the base and the proposed architecture.

Table 5.1: Comparison between classification accuracy, average ROC metrics, parameter count, and FLOPS between the baseline and the proposed architectures.

Architecture	Dataset	Classification Accuracy (%)	AUC	Parameters	FLOPS
Base	HLF	76.50	0.9429	399,901	102 M
Ultra-Low Latency		76.09	0.9396	2,605	583 k
Base	Constituents (30)	77.56	0.9473	405,643	1,676 M
Accuracy-Focused		77.81	0.9480	107,403	448 M

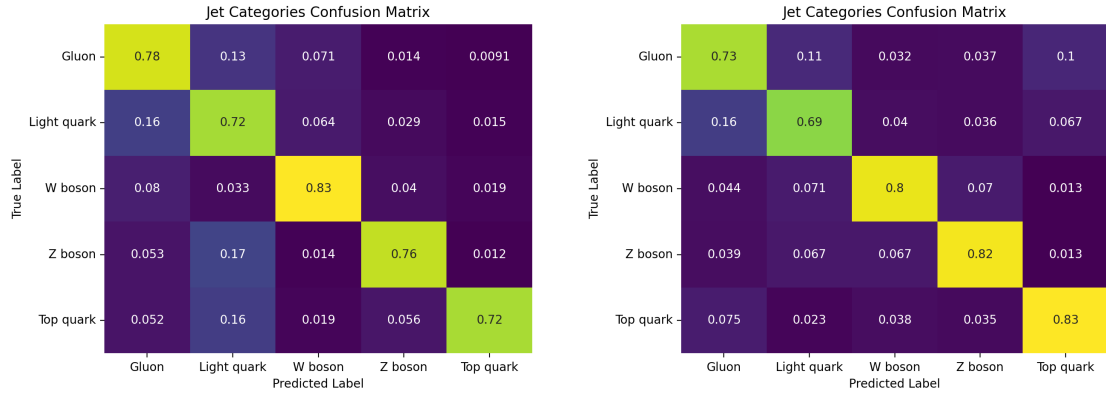


Figure 5.1: Comparison between confusion matrices of the ultra-low latency model on the HLF dataset (left) and the accuracy-focused model on the constituent list dataset (right).

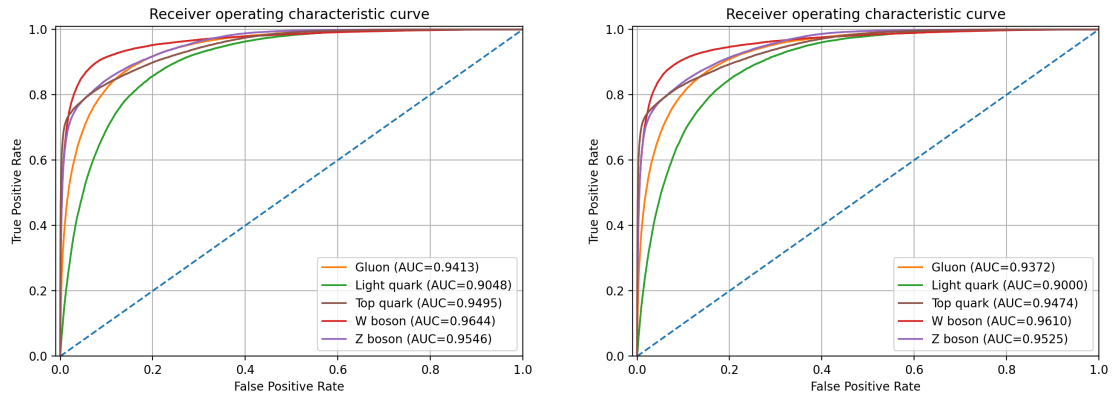


Figure 5.2: Comparison between ROC and AUC of the base (left) and ultra-low latency (right) models using the HLF dataset.

Figure 5.2 compares the base and the ultra-low latency models trained on the HLF dataset. The curves behave very similarly, with the ones associated with gluons and light quarks achieving notably lower true positive rates (TRPs) at lower false positive rates (FPRs), which means that both models have more troubles between distinguishing between these two jet categories.

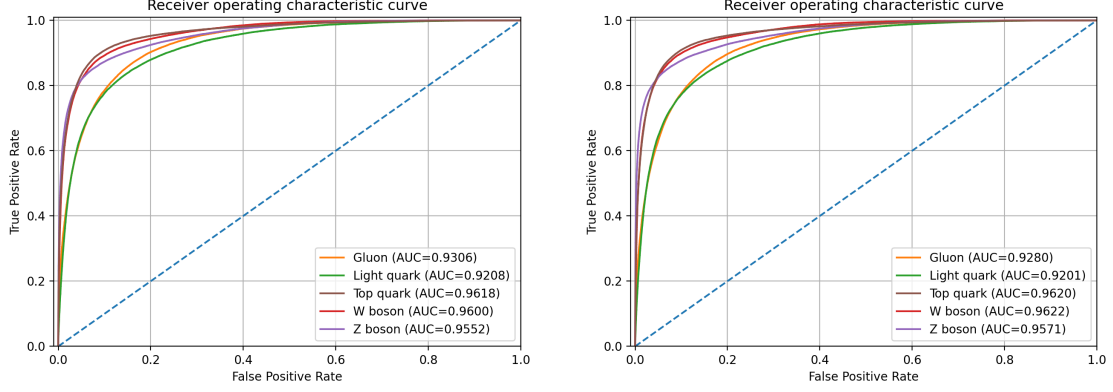


Figure 5.3: Comparison between ROC and AUC of the base (left) and accuracy-focused (right) models using the constituent list dataset.

Figure 5.3 shows an analogous comparison for the base and accuracy-focused models using the constituent list representation. Gluons and light quarks' classifications are still behind the other jets, indicating the more challenging nature of differentiating between these jet categories.

Table 5.2: AUC and true positive rates at 10 % and 1 % false positive rates for the ultra-low latency and accuracy-focused models.

Jet	Ultra-low latency			Accuracy-focused		
	AUC	TPR @ FPR		AUC	TPR @ FPR	
		10 %	1 %		10 %	1 %
Gluon	0.9372	0.8057	0.3928	0.9280	0.7765	0.3242
Light quarks	0.9000	0.6813	0.1618	0.9201	0.7729	0.2788
Top quark	0.9474	0.8304	0.7064	0.9620	0.9094	0.5553
W boson	0.9610	0.9079	0.6207	0.9622	0.8990	0.5632
Z boson	0.9525	0.8390	0.6277	0.9571	0.8768	0.6704

Table 5.2 provides a comprehensive view of the models' true positive rates at 10 % and 1 % false positive rates. Interestingly, the ultra-low latency model yields a superior classification of gluons, despite being a simpler design. This can be attributed to the HLF dataset that might capture some characteristics of gluons better than the constituent list one.

5.1.3 CPU and GPU Latency

Table 5.3 and 5.4 display the inference times for the two proposed models, measured on several CPUs and GPUs. The results serve as a baseline for quantifying the improvements thanks to the hardware acceleration, which are discussed in the next section. The experimental setup for accurately measuring the inference times involves firstly performing 5 warm-up runs, followed by CUDA-synchronization where applicable, and then timing 200 test runs, for which mean and standard deviation were calculated. It is worth pointing out that the results for the ultra-low latency model are similar regardless of the underlying hardware, including very little differences in performance between CPUs and GPUs. This can suggest that the machines are bottle-necked in terms of memory instead of computational resources, as that behavior is not observed for the more computationally-complex, accuracy-focused model.

The detailed specifications of the four machines used for measuring the inference time are listed below. All machines run CentOS 7.0 operating system, while the first listed machine, which is

Table 5.3: Comparison of ultra-low latency model’s inference times, with batch size of 128

	Device	Inference time	
		per batch (ms)	per sample (μ s)
CPU	Intel Xeon Silver 4110	1.741 ± 0.027	13.604 ± 0.207
	Intel Xeon X5690	1.622 ± 0.026	12.670 ± 0.206
	Intel Xeon E5-2620 v3	1.325 ± 0.123	10.350 ± 0.963
	Intel Xeon Gold 6154	1.167 ± 0.066	9.112 ± 0.516
GPU	Nvidia GTX 1080 Ti	1.166 ± 0.112	9.111 ± 0.876
	Nvidia TITAN X	1.154 ± 0.119	9.017 ± 0.928
	Nvidia TITAN Xp	1.062 ± 0.036	8.296 ± 0.283

the GPU host, uses CUDA version 11.5, with driver version 495.29.05. In the subsequent sections, when no comparisons between machines are discussed, the used environment is the last of the listed systems.

- Dual Intel Xeon Silver 4110 at 2.10GHz with 192 GB DDR4 RAM at 2666 MT/s,
- Dual Intel Xeon X5690 at 3.47GHz with 96 GB DDR3 RAM at 1333 MT/s,
- Intel Xeon E5-2620 v3 at 2.40GHz with 192 GB DDR4 RAM at 2133 MT/s,
- Dual Intel Xeon Gold 6154 CPU at 3.00GHz with 768 GB DDR4 RAM at 2666 MT/s,

Table 5.4: Comparison of accuracy-focused model’s inference times, with batch size of 128

	Device	Inference time	
		per batch (ms)	per sample (μ s)
CPU	Intel Xeon Silver 4110	13.681 ± 0.087	106.881 ± 0.678
	Intel Xeon X5690	16.874 ± 0.237	131.832 ± 1.851
	Intel Xeon E5-2620 v3	15.311 ± 0.444	119.616 ± 3.468
	Intel Xeon Gold 6154	7.436 ± 0.498	58.092 ± 3.888
GPU	Nvidia GTX 1080 Ti	2.941 ± 0.186	22.975 ± 1.452
	Nvidia TITAN X	3.258 ± 0.358	25.456 ± 2.797
	Nvidia TITAN Xp	2.993 ± 0.269	23.385 ± 2.099

5.1.4 Comparison with Existing Solutions

It is essential to compare the proposed architectures with recent state-of-the-art designs mentioned in chapter 2. Table 5.5 shows the key characteristics of each neural network architecture. Although the inference times were measured on different hardware platforms, it can be safely assumed that the overall trends are maintained given the results obtained in the previous subsection.

It is important to mention that the accuracy and AUC values claimed in the literature are obtained thanks to an extensive hyperparameter tuning, which is beyond the scope of this project. Nonetheless, the proposed architectures offer an interesting trade-off between the quality of results and design complexity, which is highlighted by the low inference time.

5.2 Hardware Acceleration

This section covers the results of the hardware implementation of the two proposed architectures. It also discusses the design space exploration performed using C simulation and synthesis processes available in Vivado HLS. Thanks to its high-performance, Xilinx Virtex Ultrascale+ XCU250 (variant figd2104-2L-e) was chosen as the target FPGA platform. It contains 5,376 18 Kb BRAMs, 12,288 DSP slices, 3,456,000 registers (also referred to as FFs), and 1,728,000 LUTs.

Table 5.5: Summary of networks’ inference time, accuracy, Floating-Point Operations Per Second and parameter number for optimal batch sizes, with best values in bold. Inference times were measured using ^{*}NVIDIA GTX 1080, [†]16-thread virtualized cloud CPU, [‡]NVIDIA GTX 1080 Ti.

Neural network	Inference per batch (ms)	Accuracy / aver. AUC	FLOPS	Parameters
DNN [6]	1.0 ± 0.2[*]	0.760 / 0.941	27 k	14,725
CNN [6]	57.1 ± 0.5 [*]	0.740 / 0.911	400 k	205,525
GRU [6]	23.2 ± 0.6 [*]	0.750 / 0.912	46 k	15,575
JEDI-net [6]	121.2 ± 0.4 [*]	TODO / 0.959	116 M	33,625
JEDI-net with $\sum O$ [6]	402.0 ± 1.0 [*]	TODO / 0.957	458 M	8,767
ConstituentNet-Base [8]	~133.7 [†]	0.818 / 0.966	1,553 M	289,000
ConstituentNet-Tiny [8]	~12.4 [†]	0.805 / 0.960	13 M	8,533
Ultra-Low Latency	1.2 ± 0.1 [‡]	0.761 / 0.940	583 k	2,605
Accuracy-Focused	2.9 ± 0.2 [‡]	0.778 / 0.948	448 M	107,403

5.2.1 Ultra-Low Latency Model

The first of the proposed models undergoes a series of optimization steps, with the aim of finding the fastest, yet still accurate design configuration. The steps are described below, with their impact on latency and resource utilization depicted in figure 5.4.

1. Initial implementation of the architecture.
2. SiLU activations are replaced by ReLU, leading to a slight latency improvement with a 6 % decrease in BRAMs with no sigmoid functions needing precomputing.
3. Pipelining is introduced, latency decreases nearly 1000 times, 5 times less BRAMs are needed thanks to array partitioning, but at the same time more DSPs, LUTs and FFs are instantiated to handle the parallelism.
4. Dependencies in linear layers are simplified, which decreases all resources’ utilization, but worsens latency.
5. Separate instances of linear layers without bias calculation are used, which reduces LUTs and FFs usage, but surprisingly causes a slight increase in latency.
6. More arrays are partitioned and some are explicitly set to LUTs instead of BRAMs, halving the number of the latter and reducing latency by over 40 %.
7. The optimized log softmax block replaces the naive implementation, which causes an 86 % latency improvement with marginal resource utilization decrease.
8. The *queries*, *keys*, and *values* fully-connected block is split into separate layers, reducing BRAMs thanks to more optimal weights allocation and decreasing latency.
9. Tensor product is reorganized and simplified, nearly halving latency and BRAMs, while notably reducing LUTs and FFs.
10. Tensor product gets fully unrolled, which again halves the latency, but causes a noticeable increase in overall resource utilization.
11. Top level module variables are reorganized to reduce FFs and LUTs usage.
12. The uniform design bit-width is reduced to match DSP specification to allow for a more optimal routing, halving their required number and causing an overall improvement to latency and other resources.

The final design’s resource usage can be seen in figure 5.6, while its timing results are shown in figure 5.7. Overall, the optimization process leads to an over 1,000,000 times faster inference speed, with decreases in all resources aside from LUTs.

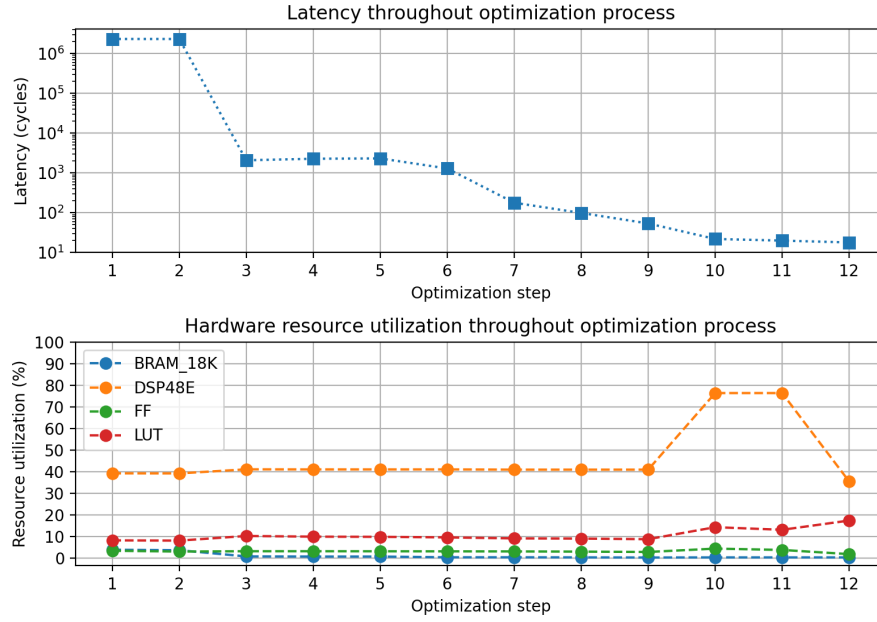


Figure 5.4: Results of the optimization process for the ultra-low latency model.

Table 5.6: FPGA resources utilization for the ultra-low latency model.

	BRAM 18K	DSP48E	FF	LUT
Total used	12	4,351	58,942	298,881
Available	5,376	12,288	3,456,000	1,728,000
Utilization	0.22%	35.41%	1.71%	17.30%

Along with the FPGA clock period of 5 ns⁷ with a 12.5 % uncertainty, the final design is able to perform a successful inference in only 90 ns with adequate slack and initiation interval of 1⁸. Thus, the hardware acceleration leads to an over 500 times decrease in latency per sample compared to the CPU and GPU implementation, and it is well within the L1T 12.5 μ s latency constraint, leaving a significant headroom for more complex processing. The test accuracy is measured as 74.4 %, which is 1.7 percent point lower than the original floating-point implementation. The average AUC is slightly lower at 0.9368, with the quantization noise being considered as the main factor for the decrease in classification capability. These values are inline with other results obtained using `hls4ml` [22], however, our implemented hardware model is over 4 times faster despite not undergoing any compression, pruning or quantization, which highlights the potential of the described optimization process.

Table 5.7: Information about design’s clock, latency and initiation interval for the ultra-low latency model.

Clock (ns)	Target	5.000	Latency	cycles	18
	Estimated	4.233		ns	90
	Uncertainty	0.620	Interval		1

It is possible for the designs other than the final one to represent interesting trade-offs between latency and hardware resources. To get a better view of that, figure 5.5 shows the latency plotted against the resources. It can be seen that the design space exploration has led to two Pareto optimal designs - the final one, which offers better latency, and the one without unrolled tensor product, which requires less resources’ utilization. In general, none of them could be called strictly better than the other. However, the final design is more suitable given the ultra low-latency objective of this work.

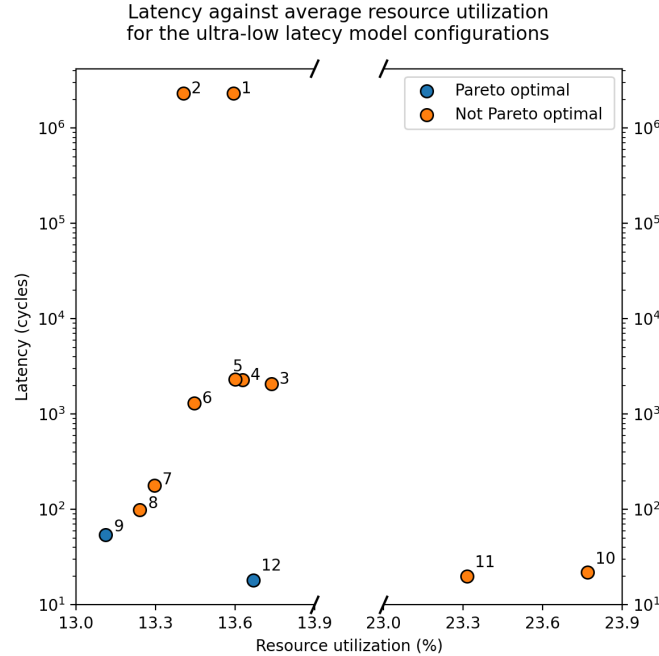


Figure 5.5: Latency plotted against average resource utilization for the ultra-low latency model configurations.

⁷5 ns period corresponds to a frequency of 200 MHz, which is used at L1T at LHC.

⁸Initiation interval informs about the number of cycles between new inputs can be processed by a pipeline. In this case, new sample can be handled each cycle.

5.2.2 Verifying Analytical Latency and DSP Models

The latency estimation formulated in the previous chapter is verified given the synthesis results - the final design requires 18 cycles as predicted. This could only be guaranteed under the assumption that every operation can be completely unrolled and pipelined and all arrays are fully partitioned, which depending on the design might require more resources than there are available. Thanks to the optimizations and simplifications used for the model, this assumption was correct, with the most demanded resource (DSPs) at around a third of its limit.

$$\# \text{ of DSPs} = L + C_{\text{linear}} \cdot (21d + 8d^2) + C_{\text{tensor}} \cdot 2 \cdot L^2 \cdot H \approx \mathcal{O}(d^2 + L^2 \cdot H) \quad (5.1)$$

The situation gets more complicated for the DSP model. After a careful analysis of the required hardware footprint of each component of the final design, it turns out that both the linear layer and the tensor product depend on more variables than just their input and output sizes, effectively making the included "constants" vary between instances. It is likely that there exists no easy-to-understand expression, given the complexity of HLS synthesis routing for pipelined designs. In other words, user-independent factors like the order in which HLS processes the design might influence the required hardware resources, including DSPs. However, there is still merit in the proposed model when it comes to its overall computational complexity, as shown in equation 5.1. While this does not provide a clear upper boundary for deciding whether a configuration exceeds the available DSP slices, it highlights the quadratic relationship with the latent dimension d , number of input jets L , as well as a linear dependency on the number of self-attention heads H .

5.2.3 Accuracy-Focused Model

The process of creating a hardware model for the accuracy-focused architecture started similarly to the previous model. After implementing the necessary layers and changing the design configuration, C simulation reported a slight decrease in test accuracy by 1.8 percent point, to a value of 76.01 % along with AUC equal to 0.944. However, very quickly it became evident that the synthesis is struggling to make any progress with the design. While it was not uncommon for it to last hours for the ultra-low latency model, this time the process was still in its early phase after days of waiting. The first solution was to keep simplifying the model to find a configuration that can be synthesized. In order to make informed decisions, a grid search was performed on the most influential design parameters, and the results can be seen in figure 5.6.

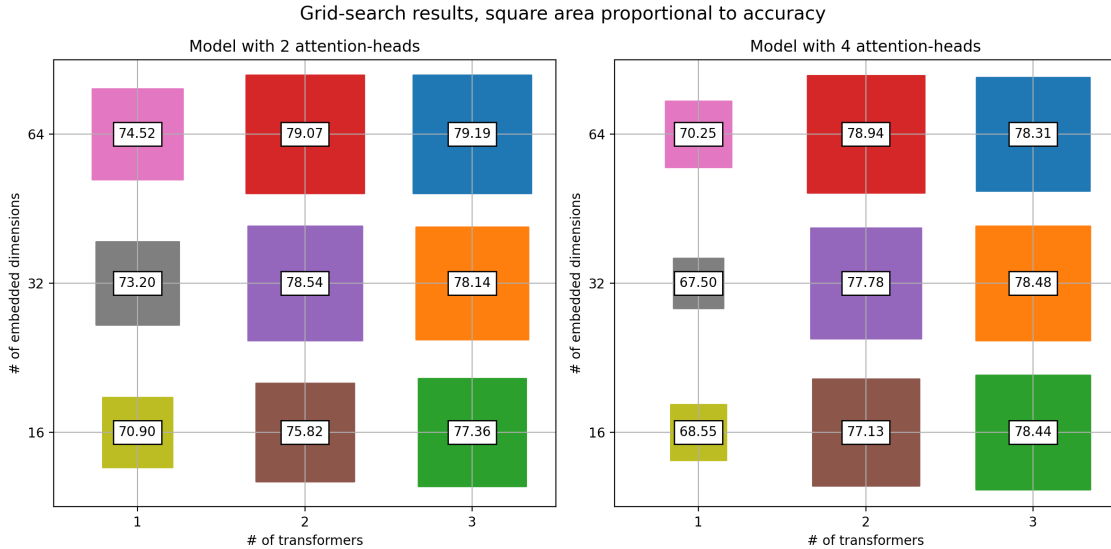


Figure 5.6: Grid-search results - squares' area proportional to accuracy.

It can be seen that reducing the number of embedded dimensions have the smallest impact, followed by the transformer layers. The model with 4 self-attention heads was evaluated in case it performed

better in certain configurations, but this concluded to not be the case. Hence, simpler designs were selected for synthesis, however, each of them would be terminated after approximately a week given that they were still in their early steps. It was clear that the design simplification alone was not enough to overcome this challenge. After careful performance analysis, it became apparent that each synthesis job keeps increasing its reserved system memory, reaching the limit at around 750 GB, which was all the server’s available RAM. This suggests that an *explosion* of the design space would happen due to the way HLS analyzes the design, and that this process was too complex to ever succeed without a fundamental change of server specifications (which is not possible given the scope of this project) or approach to synthesis. This issue is visualized in figure 5.7.

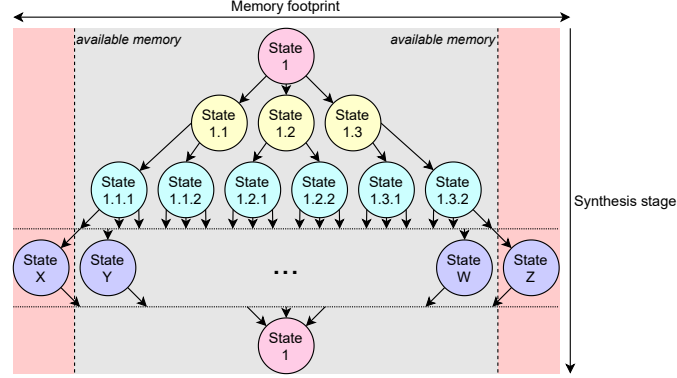


Figure 5.7: Illustration of the synthesis design space reaching the limit of available memory.

Knowing that the issue was not a matter of using a faster machine or waiting for longer, the synthesis process had to be altered to reduce the number of states generated during HLS. Thus, a serial model was created by removing pipelining and any motions of parallelizability from both the implementations created as part of this work and these already present in `hls4ml` library. Avoiding the complexity of synthesizing a pipelined design, the process finished within minutes, yielding the results shown in tables 5.8 and 5.9.

Table 5.8: FPGA resources utilization for accuracy-focused model.

	BRAM 18K	DSP48E	FF	LUT
Total used	19	15	3106	6823
Available	5,376	12,288	3,456,000	1,728,000
Utilization	0.4%	0.1%	0.1%	0.4%

It is clear that without pipelining, the model is not suitable for operating under the latency constraint of L1T at LHC. At the same time, it is interesting to note how little resources are required and how easily the synthesis process is handled for a serial design. Due to the complexity involved in the HLS, it is not feasible to interpolate the results for a pipelined design.

Table 5.9: Information about design’s clock, latency and initiation interval for the accuracy-focused model.

Clock (ns)	Target	5.000	Latency	cycles	6803
	Estimated	4.300		ns	34,015
	Uncertainty	0.620	Interval		N/A

5.3 Quantization Results

In this section, firstly the effects of quantization-aware training are explained, followed by discussion of the results generated using the proposed post-training quantization algorithm and other available

techniques. In both cases, the accuracy-focused architecture is targeted because of its increased complexity, which can highlight the impact of the quantization.

5.3.1 Quantization-Aware Training

The first analysis includes a QPyTorch implementation of the accuracy-focused architecture configured with a range of fixed-point representations. For each experiment, the same bit-width was used for every quantizable layer of the model to limit the number of configurations and focus on the high-level effects. Figure 5.8 compares several 64-bit, 32-bit and 16-bit wide floating-point and fixed-point models across epochs.

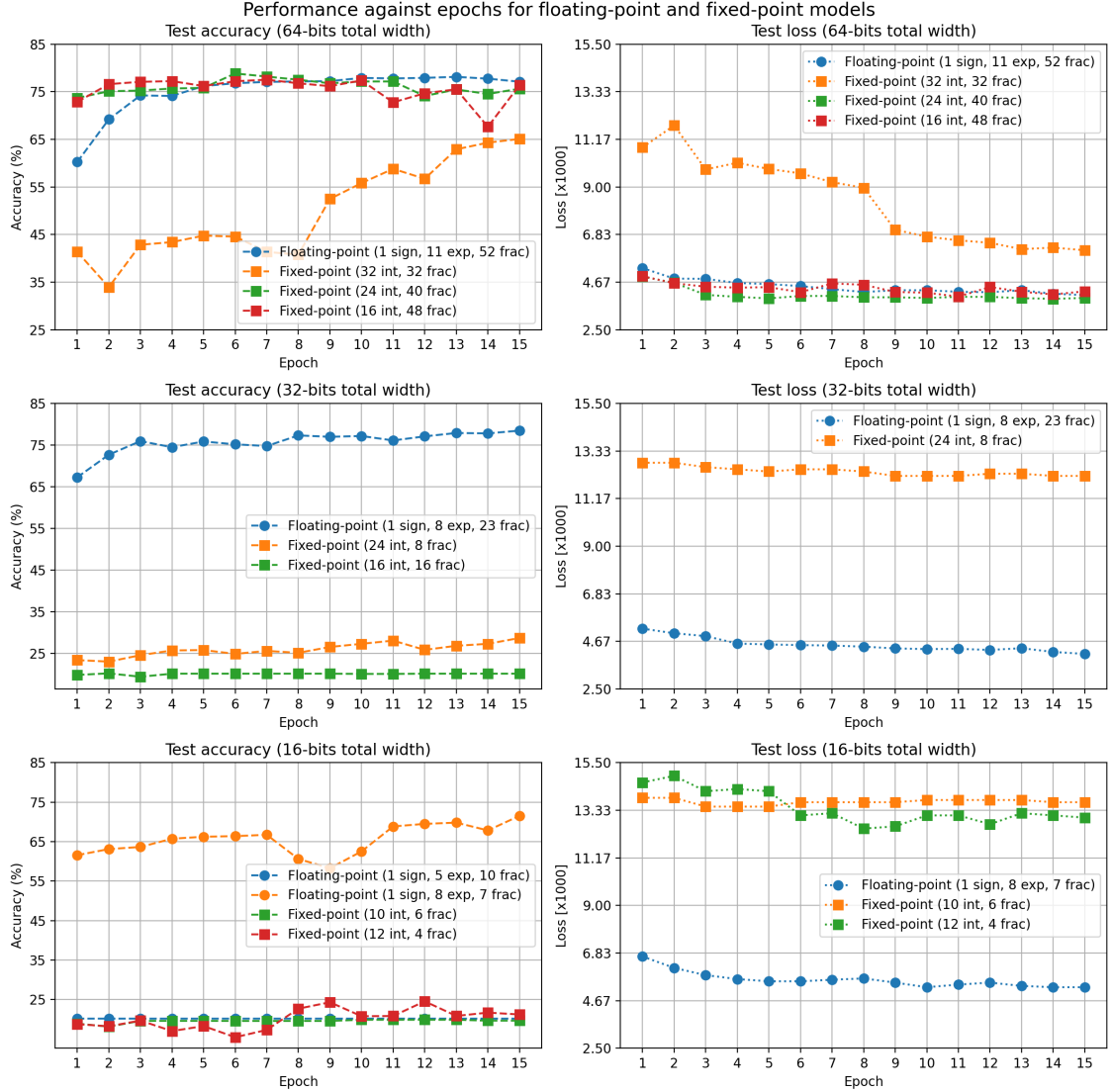


Figure 5.8: Performance against epochs for floating-point and fixed-point models.

There are two key observations that can be made. Firstly, the model appears to require a significant fractional precision for successful classification. At 64 total bits, all configurations improve with epochs, although the one with 32 bits of fractional precision takes longer to learn and ends with a worse accuracy than the rest. At 32 total bits, only the floating-point model is able to be usefully trained which further proves the impact of decimal precision. The second crucial conclusion comes from the models with 16 total bits, where only the floating-point model with 8-bit wide exponent field is able to learn, and the one other floating-point and both fixed-point models are stuck around baseline accuracy of 20 %. This suggests that the representable integer range also plays a key role

in the classification.

It is difficult to pinpoint the exact layer or mechanism involved in the model that causes the results to degrade without an adequate integer and fractional ranges. The former is predicted to be of most importance for representing the constituent list dataset features which follow wide distributions, hence value truncation leads to no useful input information which are effectively randomized without the most significant bits. The latter is believed to be most crucial for the normalization layers which map numbers to relatively small values, hence without proper decimal precision, many inputs are assigned with the same quantized outputs. Combined, these two effects mean that floating-point representations work better for this model’s quantization aware training thanks to their flexibility in precise storage of high and low values, while fixed-point numbers require both their integer and fractional fields to be relatively wide for comparable results.

The other framework mentioned in section 3.4 is Brevitas. Despite promising hardware mapping potential thanks to its close integration with hardware tool chains, at the time of this project, Brevitas lacks support for fixed-point values wider than 8 bits. It comes as no surprise given the conclusion from the previous paragraph, that this is not a wide-enough representation for this model. Although the most successful training yields a classification accuracy of only 34 %, with further development of available types, Brevitas is likely to produce comparable results to QPyTorch or potentially outperform it given its approach of offering custom quantized neural network components.

5.3.2 Post-Training Quantization

Compared to the previous quantization methods, the proposed post-training quantization approach is noticeably easier to experiment with and deploy for existing models. By simply specifying the target environment, the algorithm adapts the hardware implementation and finds the optimal set of bit-widths. In this project, 60 individual variables are analyzed for a maximum quantization granularity, which significantly differs from the quantization-aware training experiments. Figure 5.9 showcases the final integer and fractional bit-widths, with types related to the same layer logically grouped.

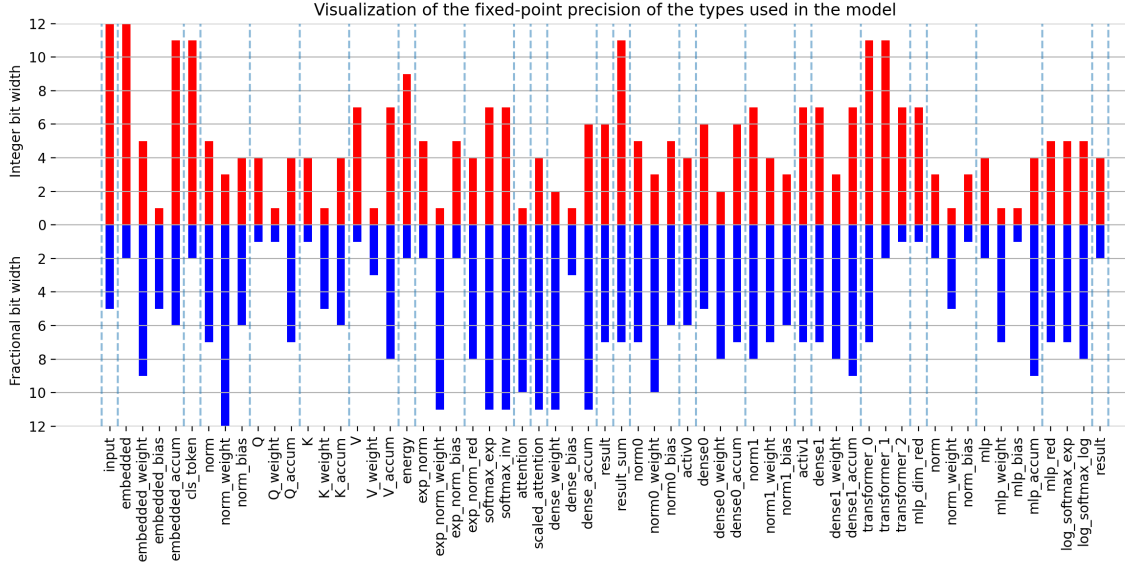


Figure 5.9: Visualization of the optimal fixed-point precision of the types used in the quantized accuracy-focused model.

This method achieved a 64 % reduction in total bit-width with a 2 % decrease to accuracy. Total bit-width relatively closely corresponds to the hardware resource utilization - while the FFs and LUTs are expected to vary linearly, DSPs and BRAMs are only affected after their width threshold is exceeded. In other words, the target FPGA’s DSP slices allow up to 34 bits as their main input,

hence only 35 bits or more lead to an additional unit being instantiated. For that reason, the algorithm uses C simulation accuracy and bit-widths as its two driving signals, without the need of performing the synthesis. By exploiting this simplified bit-width to hardware resources relation, the total run-time of the tool is significantly shorter as each simulation runs within minutes instead of hours or even not finishing at all as it was discussed before. However, the total number of simulations still lead to a program run-time of around 7 days given the model’s complexity. As a result, only one configuration of the algorithm’s positive and negative accuracy tolerances was tested, where the negative one was assigned twice the importance of the positive tolerance, driving the substantial bit-width reduction while sacrificing a fraction of the accuracy.

The success of this method proves the hypothesis about strong correlation between neighboring layers. More specifically, the average total bit-width difference between layers is measured at approximately 4 bits. Furthermore, corresponding types also seem to follow the same trends among different layers - a good example are the widths of *queries*, *keys*, and *values* linear layers, all of which have similar input, weight, and accumulator widths, with the *values*’ being slightly less similar, as they are used further along the computations than the other two. It is also worth mentioning that there is no noticeable difference in results regardless of whether the integer or the fractional parts are analyzed first.

Chapter 6

Conclusion

This work provides important insights into reconfigurable acceleration of transformer neural networks for ultra-low latency applications like high energy physics. The proposed solutions cover the whole development spectrum, from software architecture design, through various optimization techniques, to efficient hardware mapping of the neural network models. In this chapter, the main technical contributions are firstly summarized, followed by a discussion covering the existing limitations discovered during our research. Finally, several possible extensions that could improve upon our work are covered.

6.1 Achievements

This project contributes to the following research areas:

- **Hardware-Aware Neural Network Design:** We explore two neural network architectures, which combine ideas from state-of-the-art solutions and transform them into components that can be efficiently mapped to FPGAs while offering cutting-edge accuracy. The analyzed aspects involve weight and bias fusing for normalization, splitting of fully-connected layers, and preventing numerical instability in softmax activation function. To allow for an easier reasoning about the architectures, analytical models covering latency and DSP slices were developed.
- **Efficient Hardware Mapping:** We propose novel hardware blocks for tensor multiplication using Einstein Notation as well as latency and resource optimized log softmax and SiLU activations. On top of that, we further expand `hls4ml` library with efficient and customizable self-attention and transformer layers. We also introduce improvements to the fully-connected layer and the unit responsible for storing results of arbitrary precomputed functions. The resulting hardware design achieves nanosecond latency with classification quality comparable to GPU and CPU implementations that are run around thousand times slower.
- **Quantization-Aware Training:** We analyze existing solutions for quantization-aware training and adapt the most suitable framework to transformer neural networks. The conducted experiments highlight the best performing fixed-point representations and highlight potential pitfalls in very complex neural networks.
- **Post-Training Quantization:** We develop a novel algorithm for post-training quantization that can be used with existing models and target any hardware platform. Thanks to its user-defined parameters, the search can be configured to prioritize accuracy over resource utilization or vice-versa. Evaluation on the proposed models results in a nearly $2/3$ reduction of total bit-widths with negligible accuracy decrease.

6.2 Limitations

Due to the scope of the project, there was no extensive hyperparameter search involved in any of the neural network architectures. Further tuning might allow for better results, which can have a

cascading effect on the rest of this report’s analysis because the researched areas are tightly coupled. In other words, more optimized software models lead to better performing hardware implementations, which can be further amplified by carefully quantizing them during or after training. On that topic, the developed post-training quantization algorithm assumes high correlation between subsequent neural network layers. Although this was proven to be the case in our analysis, it might not generalize to every network type.

The more complex of the discussed transformer architectures was not able to be synthesized as a pipelined design using HLS due to the inherent intricacy of generating pipelined models. The serial implementation offers an interesting trade-off of a design that runs slower but requires substantially less resources. However, many real-life applications operate under strict latency constraints, which could not be met due to the existing limitations of the HLS process and the fundamental issue of transformer’s complexity.

6.3 Future Work

There are four main areas that our work could be expanded upon, which involve modifying the transformer layer, especially the self-attention mechanism, enhancing the post-training quantization method, adapting existing High-Level Synthesis optimization tools, and experimenting with other datasets.

6.3.1 Alternative Transformer Design

There is an ongoing research into alternative transformer architectures, which improve upon the quadratic complexity of self-attention, that has proven to be the main challenge of accelerating transformers. Some solutions involve low-rank approximations, sparsity, memorization, or various other techniques, summarized in figure 6.1 [65]. To tackle the complexity problem, a recent work also goes as far as exploring completely removing the self-attention layer [66].

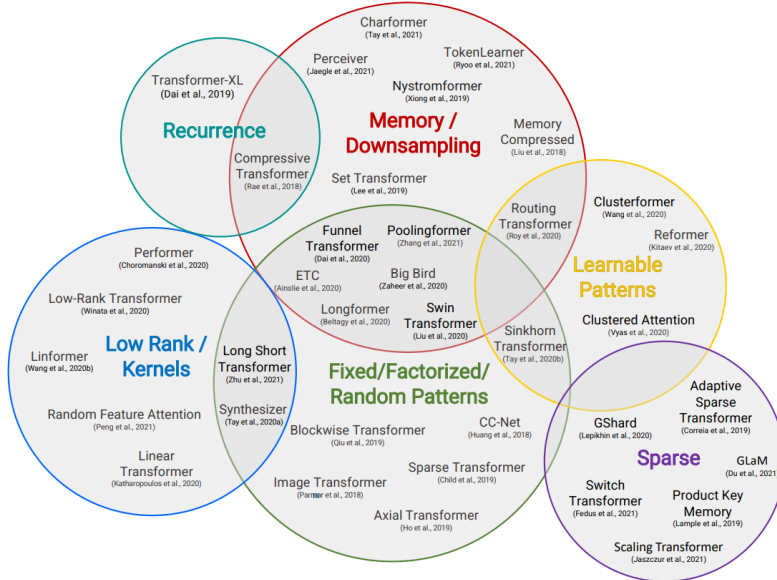


Figure 6.1: Landscape of efficient transformer architectures.

Aside from changes to self-attention, trained transformer models could be compressed using techniques like pruning to decrease the number of stored weights which can improve the inference speed and reduce hardware utilization. Additionally, the already optimized batch normalization layers could be further improved by fusing with linear layers to reduce the overall complexity. Finally, recent automatic methods of finding efficient transformer configurations [67, 68] could help with the hardware-aware architecture search challenge.

6.3.2 Post-Training Quantization Advancements

The proposed post-training quantization algorithm could base its resource estimations on an analytical model to achieve more accurate predictions without compromising the search time. A new metric of energy-consumption could also be taken into consideration to accommodate mobile or embedded platforms' limitations. In order to accelerate the search process while potentially improving its findings, the approach could also benefit from a faster simulation and awareness of tested configurations' positions with regard to the underlying Pareto front or Roofline Model.

6.3.3 Automatic High-Level-Synthesis Optimizations

A recent High-Level Synthesis framework called ScaleHLS [69] promises to generate efficient RTL designs from HLS C++ or directly from PyTorch thanks to the use of optimizations at different levels of abstraction of the underlying Multi-Level Intermediate Representation (MLIR) [70] used for PyTorch [71]. This tool was explored during the project, but similarly to the discussed quantization-aware training methods, it does not support the self-attention layer in case of the direct PyTorch model optimization. As for trying it on the already developed HLS implementation, the size and the deep hierarchy involved in this work exceeds the complexity that ScaleHLS can handle. Nonetheless, further improvements might allow this framework to reduce the development time between designing a transformer architecture using `hls4ml` and testing its hardware implementation, supporting the objective of this work.

6.3.4 Other Datasets

Both of the datasets used in this project involve classification of five jet categories. However, other examples in the literature often discuss problems with fewer categories, for instance top-tagging, which is an approach to identifying top quarks. Simpler problems are likely to require less complex architectures - an area of the design space that this work proves to be optimal for transformers.

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](https://link.springer.com/article/10.1007/JHEP02(2015)118).
- [3] Luke de Oliveira, Michael Kagan, Lester Mackey, Benjamin Nachman, et al. 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](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, et al. 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, et al. 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, et al. 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. Technical report, -09-22 2021.
- [9] K. Krasnov and R. Percacci. Gravity and unification: a review. *Classical and quantum gravity*, 35(14):143001, Jun 14, 2018. doi: 10.1088/1361-6382/aac58d. URL <https://iopscience.iop.org/article/10.1088/1361-6382/aac58d>.
- [10] J. Walz, P. Grandemange, B. Vallage, et al. The GBAR antimatter gravity experiment. Technical Report 233, Springer International Publishing, 2015. URL <http://cds.cern.ch/record/2055685>.
- [11] D. Pagano, M. Caccia, J. Fesel, S. Gerber, et al. Gravity and antimatter: the AEGIS experiment at CERN. Technical Report 1342, IOP Publishing, 2020. URL <http://cds.cern.ch/record/2714100>.
- [12] Brian Greene. How the higgs boson was found, July 2013. URL <https://www.smithsonianmag.com/science-nature/how-the-higgs-boson-was-found-4723520/>.
- [13] Trigger, DAQ and FPGAs. URL <http://www.hep.ph.imperial.ac.uk/~tapper/lecture/trigger.pdf>.

- [14] Sean Keane. CERN’s large hadron collider restarts after three-year upgrade. URL <https://www.cnet.com/science/cerns-large-hadron-collider-restarts-after-three-year-upgrade/>.
- [15] Alex Tapper. Triggering at collider experiments. URL <http://www.hep.ph.imperial.ac.uk/~tapper/lecture/CMSIndia-2020.pdf>.
- [16] E. Coleman, M. Freytsis, A. Hinzmann, M. Narain, et al. The importance of calorimetry for highly-boosted jet substructure. *Journal of instrumentation*, 13(1):T01003, Jan 9, 2018. doi: 10.1088/1748-0221/13/01/T01003. URL <https://search.proquest.com/docview/2365693051>.
- [17] Maurizio Pierini, Javier Mauricio Duarte, Nhan Tran, and Marat Freytsis. HLS4ML LHC jet dataset (30 particles), . URL <https://doi.org/10.5281/zenodo.3601436>.
- [18] Maurizio Pierini, Javier Mauricio Duarte, Nhan Tran, and Marat Freytsis. HLS4ML LHC jet dataset (50 particles), . URL <https://doi.org/10.5281/zenodo.3601443>.
- [19] Maurizio Pierini, Javier Mauricio Duarte, Nhan Tran, and Marat Freytsis. HLS4ML LHC jet dataset (100 particles), . URL <https://doi.org/10.5281/zenodo.3602254>.
- [20] Maurizio Pierini, Javier Mauricio Duarte, Nhan Tran, and Marat Freytsis. HLS4ML LHC jet dataset (150 particles), . URL <https://doi.org/10.5281/zenodo.3602260>.
- [21] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *The journal of high energy physics*, 2008:063, Apr 1, 2008. doi: 10.1088/1126-6708/2008/04/063. URL <http://iopscience.iop.org/1126-6708/2008/04/063>.
- [22] Edward Kreinar, Jennifer Ngadiuba, Zhenbin Wu, Philip Harris, et al. Fast inference of deep neural networks in FPGAs for particle physics. Technical Report 13, Institute of Physics (IOP), Apr 16, 2018. URL <http://cds.cern.ch/record/2316331>.
- [23] G. Valentino, R. W. Assmann, R. Bruce, and N. Sammut. Classification of LHC beam loss spikes using support vector machines. pages 355–358. IEEE, Jan 2012. doi: 10.1109/SAMI.2012.6208988. URL <https://ieeexplore.ieee.org/document/6208988>.
- [24] Tianqi Chen and Tong He. Higgs Boson Discovery with Boosted Trees. In Glen Cowan, Cecile Germain, Isabelle Guyon, Balázs Kegl, and David Rousseau, editors, *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, volume 42 of *Proceedings of Machine Learning Research*, pages 69–80, Montreal, Canada, 13 Dec 2015. PMLR. URL <https://proceedings.mlr.press/v42/chen14.html>.
- [25] Andrzej Skoczen, Maciej Wielgosz, and Matej Mertik. Using LSTM recurrent neural networks for monitoring the LHC superconducting magnets. Technical Report 867, Elsevier B.V, Nov 18, 2016. URL <http://cds.cern.ch/record/2234465>.
- [26] Jie Ren, Lei Wu, and Jin Min Yang. Unveiling CP property of top-higgs coupling with graph neural networks at the LHC. *Physics letters. B*, 802:135198, Mar 10, 2020. doi: 10.1016/j.physletb.2020.135198. URL <https://dx.doi.org/10.1016/j.physletb.2020.135198>.
- [27] Edward Kreinar, Zhenbin Wu, Gianluca Cerminara, Kinga Wozniak, et al. Distance-weighted graph neural networks on FPGAs for real-time particle reconstruction in high energy physics. Technical Report 3, Frontiers Media S.A, 2020. URL <http://cds.cern.ch/record/2728798>.
- [28] Abdelrahman Elabd, Vesal Razavimaleki, Shi-Yu Huang, Javier Duarte, et al. Graph neural networks for charged particle tracking on FPGAs. Dec 3, 2021. URL <https://arxiv.org/abs/2112.02048>.
- [29] Gil Keren and Bjorn Schuller. Convolutional RNN: An enhanced model for extracting features from sequential data. pages 3412–3419. IEEE, Jul 2016. doi: 10.1109/IJCNN.2016.7727636. URL <https://ieeexplore.ieee.org/document/7727636>.
- [30] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, et al. Attention-based models for speech recognition. Jun 24, 2015. URL <https://arxiv.org/abs/1506.07503>.
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, et al. Attention is all you need. Jun 12, 2017. URL <https://arxiv.org/abs/1706.03762>.

- [32] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416, 2018. URL <http://arxiv.org/abs/1803.07416>.
- [33] Jay Alammam. The illustrated transformer. URL <https://jalammar.github.io/illustrated-transformer/>.
- [34] Mohammadreza Najafi, Kaiwen Zhang, Mohammad Sadoghi, and Hans-Arno Jacobsen. Hardware acceleration landscape for distributed real-time analytics: Virtues and limitations. pages 1938–1948. IEEE, Jun 2017. ISBN 1063-6927. doi: 10.1109/ICDCS.2017.194. URL <https://ieeexplore.ieee.org/document/7980135>.
- [35] Dagli and Lammers. Possible applications of neural networks in manufacturing. page 605 vol.2. IEEE TAB Neural Network Committee, 1989. doi: 10.1109/IJCNN.1989.118423. URL <https://ieeexplore.ieee.org/document/118423>.
- [36] Cathy Wu, Michael Berry, Sailaja Shivakumar, and Jerry McLarty. Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. *Machine learning*, 21(1):177, Oct 1, 1995. doi: 10.1023/A:1022677900508.
- [37] Vincent Vanhoucke, Andrew Senior, and Mark Z. Mao. Improving the speed of neural networks on CPUs. In *Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011*, 2011.
- [38] Sainathan Ganesh Iyer and Anurag Dipakumar Pawar. GPU and CPU accelerated mining of cryptocurrencies and their financial analysis. pages 599–604. IEEE, Aug 2018. doi: 10.1109/I-SMAC.2018.8653733. URL <https://ieeexplore.ieee.org/document/8653733>.
- [39] Gang Chen, Haitao Meng, Yucheng Liang, and Kai Huang. GPU-accelerated real-time stereo estimation with binary neural network. *IEEE transactions on parallel and distributed systems*, 31(12):2896–2907, Dec 1, 2020. doi: 10.1109/TPDS.2020.3006238. URL <https://ieeexplore.ieee.org/document/9130887>.
- [40] Qianru Zhang, Meng Zhang, Tinghuan Chen, Zhifei Sun, et al. Recent advances in convolutional neural network acceleration. *Neurocomputing (Amsterdam)*, 323:37–51, Jan 5, 2019. doi: 10.1016/j.neucom.2018.09.038. URL <https://dx.doi.org/10.1016/j.neucom.2018.09.038>.
- [41] Graphcore. Graphcore intelligence processing unit. URL <https://www.graphcore.ai/products/ipu>.
- [42] Phil Knag, Jung Kuk Kim, Thomas Chen, and Zhengya Zhang. A sparse coding neural network ASIC with on-chip learning for feature extraction and encoding. *IEEE journal of solid-state circuits*, 50(4):1070–1079, Apr 2015. doi: 10.1109/JSSC.2014.2386892. URL <https://ieeexplore.ieee.org/document/7015626>.
- [43] K. Venkata Ramanaiah and Cyril Prasanna Raj. ASIC implementation of neural network based image compression. *International Journal of Computer Theory and Engineering*, pages 494–498, 2011. doi: 10.7763/IJCTE.2011.V3.356.
- [44] Andrew Boutros, Sadegh Yazdanshenas, and Vaughn Betz. You cannot improve what you do not measure. *ACM transactions on reconfigurable technology and systems*, 11(3):1–23, Dec 22, 2018. doi: 10.1145/3242898. URL <http://dl.acm.org/citation.cfm?id=3242898>.
- [45] Eriko Nurvitadhi, Ganesh Venkatesh, Jaewoong Sim, Debbie Marr, et al. Can FPGAs beat GPUs in accelerating next-generation deep neural networks? *FPGA '17*, pages 5–14. ACM, Feb 22, 2017. doi: 10.1145/3020078.3021740. URL <http://dl.acm.org/citation.cfm?id=3021740>.
- [46] Yixing Li, Zichuan Liu, Kai Xu, Hao Yu, et al. A GPU-outperforming FPGA accelerator architecture for binary convolutional neural networks. *ACM journal on emerging technologies in computing systems*, 14(2):1–16, Jul 27, 2018. doi: 10.1145/3154839. URL <http://dl.acm.org/citation.cfm?id=3154839>.

- [47] Eriko Nurvitadhi, David Sheffield, Jaewoong Sim, Asit Mishra, et al. Accelerating binarized neural networks: Comparison of FPGA, CPU, GPU, and ASIC. pages 77–84. IEEE, Dec 2016. doi: 10.1109/FPT.2016.7929192. URL <https://ieeexplore.ieee.org/document/7929192>.
- [48] Eriko Nurvitadhi, Jaewoong Sim, David Sheffield, Asit Mishra, et al. Accelerating recurrent neural networks in analytics servers: Comparison of FPGA, CPU, GPU, and ASIC. pages 1–4. EPFL, Aug 2016. doi: 10.1109/FPL.2016.7577314. URL <https://ieeexplore.ieee.org/document/7577314>.
- [49] Sakari Lahti, Panu Sjoval, Jarno Vanne, and Timo D. Hamalainen. Are we there yet? a study on the state of high-level synthesis. *IEEE transactions on computer-aided design of integrated circuits and systems*, 38(5):898–911, May 2019. doi: 10.1109/TCAD.2018.2834439. URL <https://ieeexplore.ieee.org/document/8356004>.
- [50] CERN. Facts and figures about the LHC | CERN, . URL <https://home.cern/resources/faqs/facts-and-figures-about-lhc>.
- [51] 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>.
- [52] 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.
- [53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778. IEEE, Jun 2016. doi: 10.1109/CVPR.2016.90. URL <https://ieeexplore.ieee.org/document/7780459>.
- [54] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. Feb 23, 2016. URL <https://arxiv.org/abs/1602.07261>.
- [55] Hongxiang Fan, Shuanglong Liu, Martin Ferianc, Ho-Cheung Ng, et al. A real-time object detection accelerator with compressed SSDLite on FPGA. pages 14–21. IEEE, Dec 2018. doi: 10.1109/FPT.2018.00014. URL <https://ieeexplore.ieee.org/document/8742299>.
- [56] Patrick Schaumont. Fixed point arithmetic in DSP — real time digital signal processing. URL <https://schaumont.dyn.wpi.edu/ece4703b21/lecture6.html>.
- [57] Raghuraman Krishnamoorthi, Seth Weidman, and Jerry Zhang. Static quantization with eager mode in pytorch. URL https://pytorch.org/tutorials/advanced/static_quantization_tutorial.html.
- [58] Abhishek Sharma. Pytorch JIT and torchscript. URL <https://towardsdatascience.com/pytorch-jit-and-torchscript-c2a77bac0fff>.
- [59] Jerry Zhang. FX graph mode quantization user guide. URL https://pytorch.org/tutorials/prototype/fx_graph_mode_quant_guide.html.
- [60] Alessandro Pappalardo. Xilinx/brevitas, 2021. URL <https://doi.org/10.5281/zenodo.3333552>.
- [61] Tianyi Zhang, Zhiqiu Lin, Guandao Yang, et al. QPytorch: A low-precision arithmetic simulation framework, 2019.
- [62] Alan H. Barr. The einstein summation notation. *An Introduction to Physically Based Modeling (Course Notes 19)*, pages E, 1:57, 1991.
- [63] Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurate computation of the log-sum-exp and softmax functions. *arXiv preprint arXiv:1909.03469*, 2019.

- [64] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, et al. HAQ: Hardware-aware automated quantization with mixed precision. pages 8604–8612. IEEE, Jun 2019. doi: 10.1109/CVPR.2019.00881. URL <https://ieeexplore.ieee.org/document/8954415>.
- [65] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. Sep 14, 2020. URL <https://arxiv.org/abs/2009.06732>.
- [66] Vinicius Mikuni and Florencia Canelli. Point cloud transformers applied to collider physics. *Machine Learning: Science and Technology*, 2(3):35027, Jul 13, 2021. doi: 10.1088/2632-2153/ac07f6. URL <https://iopscience.iop.org/article/10.1088/2632-2153/ac07f6>.
- [67] Henry Tsai, Jayden Ooi, Chun-Sung Ferng, Hyung Won Chung, et al. Finding fast transformers: One-shot neural architecture search by component composition. Aug 15, 2020. URL <https://arxiv.org/abs/2008.06808>.
- [68] Xiu Su, Shan You, Jiyang Xie, Mingkai Zheng, et al. ViTAS: Vision transformer architecture search. Jun 25, 2021. URL <https://arxiv.org/abs/2106.13700>.
- [69] Hanchen Ye, Cong Hao, Jianyi Cheng, Hyunmin Jeong, et al. ScaleHLS: A new scalable high-level synthesis framework on multi-level intermediate representation. *arXiv preprint arXiv:2107.11673*, 2021.
- [70] Chris Lattner, Mehdi Amini, Uday Bondhugula, Albert Cohen, et al. MLIR: Scaling compiler infrastructure for domain specific computation. In *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 2–14, 2021. doi: 10.1109/CGO51591.2021.9370308.
- [71] llvm. Torch-MLIR, July 30, 2020. URL <https://github.com/llvm/torch-mlir>.

Notes

Constituent list feature distribution	11
	11
	11
	11
	11
	11
	11
	11
Batch norm vs layer norm as both are used in the architecture	14
	14
	14
	14
	14
	14
Cite or derive this complexities	15
FPGA lattice overview to visualize explain the idea behind this technology	17
	17
	17
	17
HLS to RTL flow diagram	17
	17
	17
	17
More about hls4ml	18
Difficulty: rtl > hls > python hl4ml, draw comparison with assembly	18
	18
	18
	18
	18
reference figure from physics background when it is done	25