**Thesis Statement - 21/05/2024**

Thesis:

We are quantizing the Particle Transformer model using QKeras in order to enable it to be used in real-time at the Large Hadron Collider using FPGAs.

The Particle Transformer is an existing large neural network that was designed to classify particle jets created at the Large Hadron Collider. However, this model is too large to run effectively on a real-time system like the LHC. We are quantizing this model using QKeras in order to implement it on a FPGA for faster inference that enables it for use at the edge.

Draft:

Running large neural networks on real-time edge systems is a challenging tasks because of the time, compute constraints, and the limitation of the hardware architecture.

Abstract:

Running large machine learning models on the edge at the Large Hadron Collider is a challenging task because of the limited compute available and the time constraints. Existing models are often too large and therefore take a lot of memory to process and time to process the data. One method to get around this problem is to quantize existing models and use FPGAs (as opposed to general-purpose GPUs) for faster and more specialized processing. Our project aims to quantize an existing Particle Transformer model from PyTorch to QKeras. This new quantized model can then be implemented on an FPGA using the DeepSoCFlow library. We hope to maintain similar accuracy levels but achieve faster inference time.

**Abstract - 16/05/2024**

* Problem
* Solution
* Significance

Running large machine learning models on the edge at the Large Hadron Collider is a challenging task because of the limited compute available and the time constraints. Existing models are often too large and therefore take a lot of memory to process and time to process the data. One method to get around this problem is to quantize existing models and use FPGAs (as opposed to general-purpose GPUs) for faster and more specialized processing. Our project aims to quantize an existing Particle Transformer model from PyTorch to QKeras. This new quantized model can then be implemented on an FPGA using the DeepSoCFlow library. We hope to maintain similar accuracy levels but achieve faster inference time.

Our project aims to speed up data processing and reduce the memory read/write on the hardware to decrease data transfer and

Significance:

Our translation from Pytorch to Qkeras and hardware AIP implementation could make data processing more efficient, the quantization process increased MACs, which decreased the overall run time of training large datasets using machine learning technology (transformer) while keeping the accuracy of the model classification.