Project Proposal : Fast Inference for Quantized Transformer Attention Heads

CSE 237C Project 3
Carmen Dyck A59016086
Sanjayan Sreekala A59020260

Overview:

With the upsurge in large language model (LLM) applications, the demand for efficient inference methods for transformers is pressing. Current techniques focus largely on training acceleration, yet inferencing—calculating outputs with fixed weights—is critical for serving commercial LLMs at scale and for creating quality synthetic text data for training future versions of LLMs. This project focuses on developing a method for hardware-accelerated inferencing of transformers, specifically a 4-bit quantized attention head component. By achieving this, we aim to set the stage for a complete transformer capable of 1000s of token inferences per second.

Description:

The project's goal is to implement the attention mechanism of a 4-bit quantized transformer head, which includes QKV computation, softmax approximation, and output processing through a dense layer, on a fixed token input size. The approach will explore softmax function alternatives and optimize data movement within the hardware to enable parallel computation and efficient use of DSPs for multiple 4-bit operations.

Deliverables:

- A hardware implementation of QKV computation, softmax, and dense layer processing in HLS.
- A comprehensive testbench for the hardware implementation.
- Implementation on an FPGA board.
- A final project report detailing the design, implementation, and testing process.

Timeline:

Week 1: Implementation of QKV computation with 4-bit multiplication.
Week 2: Implementation of softmax and dense layer computation.

Week 3: Optimization of data movement on-chip and off-chip.
Week 4: Testing, debugging, and report writing.

Project Requirements:

The project might necessitate a high memory FPGA. We will assess the current resource requirements during preliminary research and notify the instructor should there be a need for specialized hardware.

This project aims to leverage skills acquired in CSE 237C to demonstrate a tangible solution to the growing need for transformer model inferencing.