Take-home Assignment

Problem Description

Given a pre-trained Llama, minimize the bit-width of the data types of various parts of the model (e.g. weights, activations, etc.) while at the same time maximising the test accuracy of the model on the CoQA dataset. You are free to make any design choices about the quantization method and how it is applied, but consider the effects of the chosen quantization on the hardware performance metrics.

Constraints

You should use Llama 3.2-1B[1] and the lm-evaluation-harness to test the accuracy on CoQA[2]. Your code and analysis should be targeted for GPU but feel free to use any GPU available to you. We do not expect you to use CUDA or Triton kernels for the quantization part.

Deliverables

Report

Use the report to explain and justify your design choices, experimental setup, and observed results. The report should be based on the template of a known scientific venue. The maximum report length, excluding references, is 4 pages. There is no need to go through related work or background in detail.

Code

Structure your code so that it is modular, easy to read, and easy to run. We will use your code to recreate your results so make sure to include the hyper-parameters that you used to run the code.

[1] Pretrained checkpoint available here.

[2] Available here