Assignment 1: Scalable Transformer Training & Profiling

CS395T: Foundations of Machine Learning for Systems Researchers
Assignment 1: Scalable Transformer Training & Profiling
Due on Canvas: September 18th, 11.59pm

Objective

The goal of this assignment is to give you hands-on experience with training a modern Transformer-based model, profiling its computational performance, and applying systems-level optimizations to improve efficiency. You will investigate trade-offs between model size, sequence length, and throughput while maintaining accuracy.

Assignment Description

1. **Model and Dataset**
   - Train a small Transformer or Vision Transformer on one of the following:
     - **Language Modeling:** WikiText-103.
     - **Image Classification:** ImageNet-100 or TinyImageNet.
2. **Profiling**
   - Measure GPU memory usage and throughput scaling by comparing:
     - context length or image resolution vs. GPU memory usage
     - batch size vs. GPU memory usage
     - batch size vs. throughput
     - mixed precision on/off vs. GPU memory usage
     - mixed precision on/off vs. throughput
3. **Optimizations**
   - Implement and evaluate at least two of the following:
     - Gradient checkpointing
     - Fused operations (e.g., FlashAttention, `xformers` kernels)
     - Distributed Data Parallel (DDP) training
     - Parameter-efficient training (LoRA, QLoRA)
4. **Analysis**
   - Identify trade-offs between:
     - model size vs. task performance (accuracy)
     - context length or image resolution vs. task performance (accuracy)
   - Discuss how your chosen optimizations affected performance and accuracy.
     - ex) DDP on/off vs. throughput
     - ex) xformers on/off vs. memory usage

Deliverables

1. **Code:** Well-structured and documented code implementing your training, profiling, and optimizations.
2. **Report (3–4 pages):**
   - Experimental setup: dataset, model architecture, hardware used.

  - Profiling results with tables and plots.
  - Backgrounds on the chosen optimization details.
  - Analysis of trade-offs and key observations.

3. **Plots:**
  - GPU memory usage vs.. sequence length / image resolution.
  - Throughput vs.. batch size (with and without mixed precision).
  - Before/after optimization comparisons.

Extra Credit

Implement an additional advanced optimization technique, such as:

- Pipeline parallelism or tensor parallelism.
- Quantization-aware training (QAT).
- Asynchronous data loading with GPU-based preprocessing (e.g., NVIDIA DALI).

Include an additional section in your report analyzing the effects of this method.

Grading Breakdown

- Model training and baseline profiling: 20%.
- Implementation of optimizations: 40%.
- Report (plots, quality of analysis, discussion of trade-offs): 40%.
- Extra credit: up to +10%.