Exploring Optimizations for Inference of Deep Learning Models on Edge Device(s)

By Adejuwon Fasanya

Objective

The objective for this project would be to explore and evaluate various optimization techniques intended to make deep learning models capable of real-time (or as close as we can manage) inference on edge devices. The main target will be my personal laptop, a 2018 13-inch MacBook Pro, with possible stretch goals for other devices if time permits.

Challenges

The project will require the implementation, application and evaluation of various optimization techniques for deep learning inference. The target device itself will be a relatively old laptop with limited hardware. This hardware could present additional challenges as the frameworks we use might have limited support (e.g an 8th gen Intel CPU with integrated graphics) which might require exploration into new frameworks/techniques or even attempting to fill these gaps ourselves when reasonable.

Approach

I plan on exploring a variety of inference optimizations which may include but not be limited to

- Model Pruning
- Model Distillation
- Low Rank Factorization/Approximation
- Quantization/Mixed Precision
- torch.compile()

These optimizations will be evaluated both individually and in conjunction with one another in order to understand their impact on performance. I'd also like to use performance profiling to understand the most prominent bottlenecks for inference on my machine and how the optimization techniques alleviate these bottlenecks.

Implementation Details

Edge Device (for inference) - 2018 MacBook Pro

- CPU: 2.3 GHz Quad-Core Intel Core i5 (Gen 8)
- GPU: Intel Iris Plus Graphics 655 1536 MB
- RAM: 8 GB 2133 MHz LPDDR3

Additionally I may use a Google Cloud instance with GPU support for the purpose of teacher-student training if I explore model distillation. However optimization for this use case will not be the focus of the project.

Software: Most work is expected to be using Pytorch, but there is potential to explore exporting models to other frameworks, particularly if they provide better support for the available hardware.

Models: Currently I'd like to focus on models for image/video or audio classification as this gives a concrete target for what "real-time inference" would be (e.g something on the order of 30 predictions per second). These would probably be simpler convolutional models. I'd also consider exploration into language models but these would likely face much greater constraints from our hardware.

Datasets: Datasets would mostly be used for eval or distillation since the focus of the project is on inference. The Ravdess dataset provides both audio and visual data for emotional classification which might serve as a good target.

Demo Planned

The demo would likely take the form of a "before and after" comparison showing off the real-time performance of the unoptimized and optimized models. This would provide an opportunity to see the concrete gains that the optimizations provided.

References:

1. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
2. https://arxiv.org/pdf/2006.08129