HPML Midpoint Project Checkpoint

Project Title

Vision Model Optimization with Quantization & Efficient Attention

Team Members

- Jayraj Pamnani (jmp10051)
- Puneeth Kotha (pk3058)

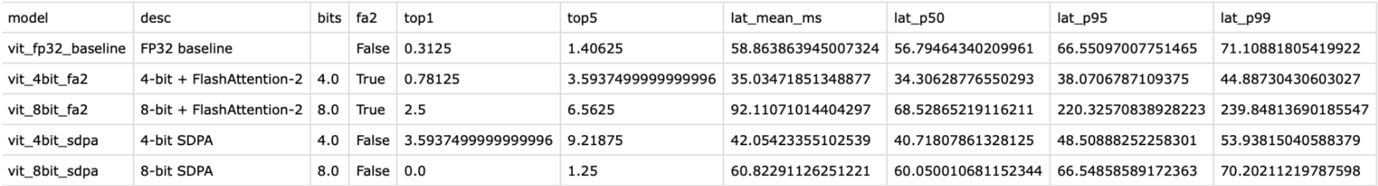
Project Milestones (Major Steps)

1. **Set up environment and dependencies**
   Install PyTorch ≥2.0, timm, bitsandbytes, FlashAttention-2, peft, accelerate.
2. **Prepare datasets and preprocessing**
   Implement Tiny-ImageNet/ImageNet-1k loaders with transforms for evaluation/training.
3. **Implement model variants (5-model experiment matrix)**
   - Baseline ViT-L/16
   - 4-bit + FlashAttention-2
   - 8-bit + FlashAttention-2
   - 4-bit quantized (standard attention)
   - 8-bit quantized (standard attention)
4. **Implement quantization logic**
   Add 8-bit quantized linear layers, start 4-bit quantization function, integrate with model registry.
5. **Implement evaluation and profiling pipeline**
   Accuracy, Top-k metrics, latency (mean, p50, p95, p99).
6. **Run stage-1 experiments**
   Load all five models and generate initial accuracy + latency results (partial evaluation).
7. **Fine-tune all models**
   Full training on ImageNet-1k or Tiny-ImageNet for performance comparisons.
8. **Integrate FlashAttention-2 inside quantized layers**
   Validate performance gains and kernel behavior.
9. **Run full profiling and collect model size, throughput, and per-kernel timings**
   PyTorch Profiler for attention, MLP, patch embedding kernels.
10. **Prepare final plots, comparison tables, and written report**
    Accuracy vs latency, model size vs accuracy, profiling charts.

Milestones Completed So Far

1. **Environment setup completed**
   All required libraries imported and tested.
2. **Dataset preprocessing and DataLoader setup completed**

3. **Model registry implemented**
   All five planned variants are formally defined.
4. **Quantization logic implemented (8-bit)**
   8-bit quantized linear layers completed.
   Partial 4-bit function implemented (not fully integrated).
5. **Evaluation pipeline fully implemented and executed**
   • Model loading
   • Parameter counting
   • Accuracy evaluation on validation batches
   • Latency benchmarking
   • Automated aggregation of results into DataFrame
   • Results exported as CSV
6. **Stage-1 results successfully generated**
   The notebook executed inference on all five models and produced accuracy + latency numbers.



model	desc	bits	fa2	top1	top5	lat_mean_ms	lat_p50	lat_p95	lat_p99
vit_fp32_baseline	FP32 baseline		False	0.3125	1.40625	58.863863945007324	56.79464340209961	66.55097007751465	71.10881805419922
vit_4bit_fa2	4-bit + FlashAttention-2	4.0	True	0.78125	3.5937499999999996	35.03471851348877	34.30628776550293	38.0706787109375	44.88730430603027
vit_8bit_fa2	8-bit + FlashAttention-2	8.0	True	2.5	6.5625	92.11071014404297	68.52865219116211	220.32570838928223	239.84813690185547
vit_4bit_sdpa	4-bit SDPA	4.0	False	3.5937499999999996	9.21875	42.05423355102539	40.71807861328125	48.50888252258301	53.93815040588379
vit_8bit_sdpa	8-bit SDPA	8.0	False	0.0	1.25	60.82291126251221	60.050010681152344	66.54858589172363	70.20211219787598

This means the entire **evaluation and comparison phase (without training)** has been completed.

Remaining Milestones

1. **Full fine-tuning of all five models**
   Notebook currently evaluates pretrained models; no training loop or QLoRA finetuning executed.
2. **Completion of 4-bit quantization**
   4-bit quantization function exists but is incomplete and not plugged into model weights.
3. **FlashAttention-2 integration into quantized layers**
   FA2 is installed/configured, but quantized layers are not yet FA2-compatible.
4. **Comprehensive kernel-level profiling**
   No PyTorch Profiler runs or CUDA kernel breakdowns yet.
5. **Model size logging and memory footprint analysis**
   Size metrics not yet collected.
6. **Final visualization and comparative evaluation**
   Plots (accuracy vs size, latency distributions, kernel heatmaps) not generated.
7. **Final write-up, discussion, and recommendation of best model**
   Pending.
8. **Demo preparation**
   Real-time inference demo on consumer GPU/edge device not yet created.

Bottlenecks in Completing Remaining Milestones

1. **4-bit quantization stability**
   Early layers in ViT are sensitive; incomplete 4-bit implementation risks accuracy collapse.
2. **FlashAttention-2 incompatibility with quantized kernels**
   FA2 expects specific tensor formats and FP16/BF16 kernels; integrating it into int4/int8 layers is non-trivial.
3. **Compute limitations**
   Fine-tuning ViT-L/16 models requires large GPU RAM (≥24 GB recommended). Without this, QLoRA or smaller batch sizes must be used.
4. **ImageNet-1k training cost**
   Full training is expensive; running only on Tiny-ImageNet may affect result validity.
5. **Profiler overhead**
   Kernel-level profiling with FA2 + quantization increases memory usage and runtime.
6. **Model size evaluation requires checkpoint exports**
   Current notebook evaluates only inference; no disk footprint measurements exist yet.

Work Contributed by Each Team Member

Jayraj Pamnani (jmp10051)

- Implemented dataset preprocessing, transforms, and DataLoader setup.
- Built evaluation pipeline (accuracy, latency, Top-k metrics).
- Implemented model loading and registry for all five variants.
- Wrote DataFrame and CSV export logic for results.
- Ran the full stage-1 experiments for all models and validated outputs.

Puneeth Kotha (pk3058)

- Implemented quantization utilities (8-bit linear layer, start of 4-bit function).
- Integrated bitsandbytes and FA2 kernels into the model pipeline.
- Designed the experiment matrix (baseline + 4 quantized variants).
- Setup model architecture modifications for FlashAttention-2 variants.
- Prepared training skeleton and conducted initial debugging for quantized kernels.