Vision Model Optimization with Quantization & Efficient Attention

Team:
Jayraj Pamnani (jmp10051)
Puneeth Kotha (pk3058)

TL;DR: Train and compare five ViT variants — a full-precision baseline plus four quantized variants (with and without FlashAttention-2) — to produce a production-ready image classifier that cuts model size and latency while keeping Top-1 accuracy within ~2% of FP32. Target deployment on consumer GPUs and edge devices.

1. What we will train (experiment matrix - five models)

1. Baseline: ViT-L/16 (full precision FP32/FP16 baseline).
2. 4-bit + FlashAttention-2: Aggressively quantized (4-bit) ViT with FlashAttention-2 integrated.
3. 8-bit + FlashAttention-2: Quantized (8-bit) ViT with FlashAttention-2 integrated.
4. 4-bit ViT (no FA2): Quantized (4-bit) ViT using standard SDPA kernel/optimized ops (no FlashAttention-2).
5. 8-bit ViT (no FA2): Quantized (8-bit) ViT using standard SDPA kernel/optimized ops (no FlashAttention-2).

We will train/finetune each variant under the same dataset, augmentation, and hyperparameter protocol so comparisons are apples-to-apples.

2. Objectives & Quantitative Targets

Primary: deliver a compact, low-latency ViT image classifier suitable for consumer GPUs/edge.
Targets:

- Top-1 accuracy within ~2–3% of FP32 baseline.
- Model size reduction 4–8× (4-bit goal).
- Inference speedup 2–3× vs baseline via quantization + attention optimizations.
- Produce a clear comparison across the five models (accuracy, latency, memory, per-kernel profiling).

3. Key Challenges

- Aggressive quantization may distort spatial features and attention maps.
- Patch embedding and early layers are sensitive to low precision.
- Integrating FlashAttention-2 across quantized kernels requires kernel/format work.
- System constraints: memory bandwidth and CPU preprocessing overhead.

4. Approach & Techniques

- Model family: ViT (ViT-L/16).
- Quantization: PTQ and QAT where needed using bitsandbytes + QLoRA-style flow for finetuning critical layers; layerwise mixed precision (protect first/last layers).

- Attention: SDPA fusion baseline; integrate FlashAttention-2 for the two FA2 variants. Profile to decide when FA2 yields net wins.
- Parameter-efficient fine-tuning: LoRA / QLoRA for low-memory finetuning of quantized models.
- Profiling: PyTorch Profiler + per-op kernel timings to measure where gains come from (patch embed, attention, MLP).

5. Implementation

- Hardware: min NVIDIA T4 (16 GB); recommend A10/A100/RTX-4090 for larger runs. CPU ≥8 cores, RAM ≥32 GB, SSD ≥200 GB.
- Software: PyTorch ≥2.0, timm/transformers, bitsandbytes, accelerate, peft (LoRA/QLoRA), datasets (HF), albumentations. Monitoring with TensorBoard / W&B.
- Datasets: ImageNet-1k primary; Tiny-ImageNet for rapid prototyping.

6. Evaluation Plan & Metrics

For each of the five models report:

- Accuracy: Top-1 / Top-5, F1 (if class balance matters).
- Latency: ms/image (mean, p50, p95, p99), throughput (img/s).
- Model size: disk & GPU footprint.
- Kernel profiling: per-op CUDA times, memory bandwidth, and where FlashAttention-2 helps or hurts.
- Ablations: layerwise quantization sensitivity and effect of keeping first/last layers higher precision.

We will present a clear comparative table and plots (accuracy vs size, accuracy vs latency).

7. Expected Outcomes & Contributions

- Practical: 3–4× smaller models (4-bit) with 2–3× inference speedups enabling consumer/edge deployment in many scenarios.
- Scientific/engineering: systematic comparison of FlashAttention-2 + quantization vs standard quantized ViTs; layer-wise quantization guidelines; kernel-level profiling insights.
- Deliverables: reproducible training scripts, benchmark suite, profiling reports, model checkpoints, and deployment notes.

8. Demo Plan (what we will show)

- Side-by-side performance comparison of all five models (accuracy, latency, size).
- Real-time inference demo on a consumer GPU or edge device highlighting speed and memory wins.
- Profiling insights: PyTorch Profiler/TensorBoard visualizations showing where FlashAttention-2 and quantization save time.
- Deployment sketch: steps & notes to deploy the preferred quantized model on edge or cloud GPU.

9. Core References

Research Papers:
[1] Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR 2021.
https://arxiv.org/abs/2010.11929

[2] Dettmers et al., *QLoRA: Efficient Finetuning of Quantized LLMs*, NeurIPS 2023.
https://arxiv.org/abs/2305.14314

[3] Dao et al., *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning*, ICLR 2024.
https://arxiv.org/abs/2307.08691