Efficient Model Quantization and Deployment on Apple Silicon with MLX

Theodore Villalva, DJ Ravenell, Ryan Caudill
Department of Computer Science
University of South Carolina

PROJECT PROPOSAL

1.1 PROBLEM

Deep learning models such as large language models (LLMs) often require substantial memory and computational resources, making them difficult to deploy on consumer-grade hardware or low-power devices. This limits their accessibility and practical usage outside of high-performance data centers. Efficient quantization and deployment pipelines can reduce model size and inference latency while maintaining accuracy, enabling real-time applications on Apple Silicon and other edge devices.

1.2 LITERATURE REVIEW

For context and background, we will examine:

1. Profiling Apple Silicon Performance for ML Training (arXiv, 2025), which benchmarks Apple Silicon hardware and provides insight into efficiency and limitations when training ML models.
2. Towards Large-scale Training on Apple Silicon (OpenReview, 2025), which explores methods for scaling deep learning training on Apple Silicon and discusses optimizations for hardware utilization.

1.3 DATA

For this project, we plan to use publicly available benchmark datasets rather than collecting new data, since the focus is on quantization, benchmarking, and deployment. Depending on scope, we may use CIFAR-10 for ResNet quantization, as it is a well-known computer vision dataset that makes it easy to evaluate accuracy versus inference speed while remaining lightweight enough for Apple Silicon.

Alternatively, we may use WikiText-2 for LLaMA-style quantization, which is a popular corpus for language modeling and highlights experience with transformer workflows. If time allows, we could incorporate both datasets to demonstrate quantization in both computer vision and NLP, giving a broader view of optimization and deployment across domains.

1.4 METHOD

We plan to implement post-training quantization in MLX to reduce model precision (e.g., int8 and int4) for ResNet and LLaMA-style models. Where possible, we will start from existing MLX implementations and adapt them to support quantization-aware inference. Our modifications will focus on benchmarking inference speed and memory efficiency on Apple Silicon, and we will package the models into a deployable inference service (e.g., via FastAPI). This approach highlights both optimization and production readiness.

1.5 EVALUATION

We will evaluate our results by comparing model accuracy before and after quantization to measure the trade-off between compression and performance. Quantitatively, we will report metrics such as classification accuracy (for CIFAR-10), perplexity (for WikiText-2), inference latency, and throughput. Qualitatively, we will present plots and figures showing accuracy versus model size and inference speed across quantization levels. These results will demonstrate the effectiveness of quantization on Apple Silicon and the engineering trade-offs involved.