Model Quantization

Quantization Granularity

Quantization is a magic spell to reduce the memory footprint of a model. But often quantization leads to a drop in the accuracy of the model. This is where the Granularity of quantization comes into the picture. Selecting the right granularity helps in maximizing the quantization without much drop in the accuracy performance

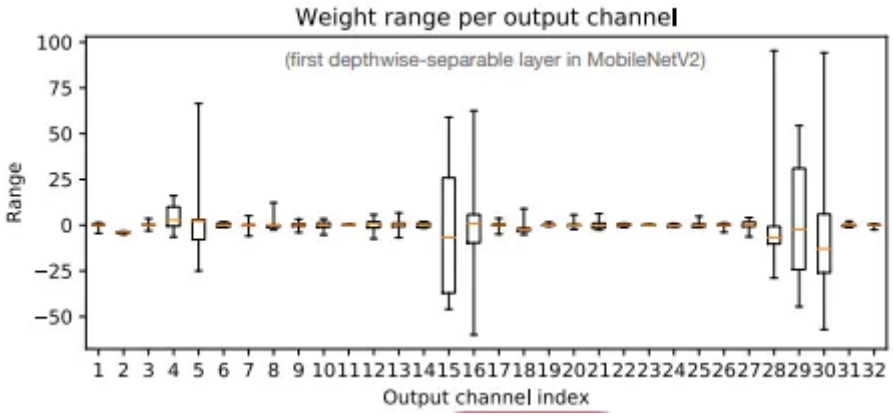
Per-Tensor Quantization

In per-tensor quantization, the same quantization parameters are applied to all elements within a tensor. Applying the same parameters across tensors will cause a drop in accuracy.

Per-Channel Quantization

In per-channel quantization, different quantization parameters are applied to each channel of a tensor independently. This often leads to a lower error while quantizing compared to per tensor quantization.

Per-channel quantization captures variations in different channels more accurately. This usually helps in CNN models where the range of weights varies over different channels.



Mix-precision Quantization

different layer using different precision.

Quantization Method

Quantization Method	Accuracy	Model Size Reduction	Inference Speed	Ease of Implementation
GGUF	High (adaptive quantization of parameter groups)	Significant (grouped quantization)	High (optimized through grouped quantization)	Moderate (requires parameter grouping)
AWQ	High (asymmetric quantization for accuracy)	Moderate	Moderate to High (asymmetric approach)	Moderate (asymmetric scaling)
GPTQ	Moderate (post-training fine-tuning)	Moderate	Moderate (benefits from post-quantization)	Easy (applied post-training)
GGML	High (mixed-precision within groups)	Significant (effective group-wise precision)	High (mixed-precision within groups)	Complex (group-wise mixed-precision)
PTQ	Moderate (quick and easy implementation)	High (straightforward implementation)	High (quick deployment)	Easy (quick application)
QAT	Very High (trained with quantization in mind)	Moderate to High (requires complex training)	High (optimized for quantization during training)	Complex (requires training with quantization)
Dynamic Quantization	Moderate (applied during inference)	High (applied at inference time)	Moderate to High (quick deployment)	Easy (inference-based)
Mixed-Precision	High (selectively applied different precisions)	Moderate to High (selectively applied)	High (selective precision boosts performance)	Complex (selective precision assignment)

Revision #4
Created 21 February 2025 03:57:54 by Mark
Updated 21 February 2025 04:35:30 by Mark