CMSC828G: Special Topics in Systems for Machine Learning

Quantization

Instructor: Yafan Huang



DEPARTMENT OF
COMPUTER SCIENCE

Why Quantization?

- Modern ML models require enormous memory capacity and bandwidth.
- Quantization reduces the number of bits used to represent numerical values.
- Potential benefits:
  - Reduced memory footprint and data movement.
  - Higher throughput with supported low-precision hardware.
  - Larger models and batches under limited GPU memory.

Quantization: Short Definition

- Quantization maps continuous or high-precision numerical values to a finite set of discrete representations.
- Quantization: x->q (e.g., 1.23 -> 24)
- Dequantization: q->x’ (e.g., 24 -> 1.20)
- Typical lossy processing: x!=x’ (e.g., 1.23!=1.20)
- Key tradeoff: Representation precision vs. storage and computational cost.

Mainstream Quantization Method

- Uniform Quantization: INT8, INT4
  - Equally spaced levels.
- Non-uniform Quantization: NF4 (from QLORA paper), Codebooks.
  - Non-equally spaced levels.
- Low-Precision FP: FP8 (e.g., E4M3, E5M2, MXFP8), FP4 (e.g., NVFP4)
  - Exponent + mantissa (or with scaling factors).

Uniform Quantization: Min-Max Scaling

- Uniform quantization uses equally spaced quantization levels.
- Min-max scaling maps a floating-point range to a fixed integer range.
- Quantization parameters:
  - Scale ***s***: Distance between adjacent levels.
  - Zero-point ***z***: Offset for representing 0.

The most popular quantization method!

$$s = \frac{x_{\max} - x_{\min}}{q_{\max} - q_{\min}}$$

$$z = \text{round}(q_{\min} - \frac{x_{\min}}{s})$$

$$q = \text{clip}(\text{round}(x/s) + z, q_{\min}, q_{\max})$$

$$x' = s(q - z)$$

Example

Original FP range: [-1.0, 3.0]

Quantized Integer range: [0, 15]

Choose scale factor:

s = (3-(-1))/(15-0) = 4/15

Choose zero offset:

z = round(0-(-1)/(4/15)) = 4

Ori. x	Int. q	Rec. x'
-1	0	-16/15≈-1.07
0	4	0
1	8	16/15≈1.07
3	15	44/15≈2.93

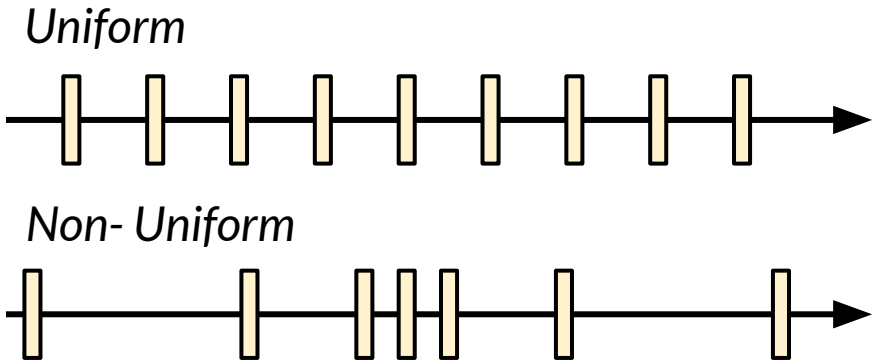
Uniform Quantization: Alternative Designs

Approach	**How is the quantization step selected?**
Min-Max	Data range + bitwidth
Symmetric Min-Max	Maximum absolute value + bitwidth
Error-bounded	Error Tolerance

$$q = \text{round}(\frac{x}{2e}),\ x' = 2eq,\ |x - x| \leq e$$

Non-Uniform Quantization

- Uniform quantization assigns equally spaced numerical levels.
- Non-uniform quantization allocates levels according to numerical distributions or other optimization objectives.
- Common approaches:
  - Logarithmic quantization.
  - Codebook-based quantization.
  - NormalFloat (NF4).



Low-Precision Floating Point

- Floating-point formats represent values using sign, exponent, and mantissa.
- Common low-precision formats:
  - FP8 E4M3: Higher precision, smaller dynamic range.
  - FP8 E5M2: Larger dynamic range, lower precision.
  - NVFP4: E2M1, 4-bit two-level block-scaled FP for Blackwell architecture.
- Floating-point quantization can also use scaling factors.
- Hardware support is critical for efficient low-precision computation.

Quantization in MLSys: Opportunities and Challenges

- Weight-only quantization
  - Reduces model memory, beneficial for memory-bound inference.
- Weight and activation quantization
  - Enables low-precision matrix multiplication on supported hardware, but requires careful control of activation quantization error.
- KV cache quantization
  - Reduces memory usage and cache traffic and improves serving concurrency.
- Where? Per-tensor, per-channel, per-group, etc.
- Lower precision does not always mean higher performance!

Student Presentations

- [NeurIPS’23] QLORA: Efficient Finetuning of Quantized LLMs
- [ICLR’23] GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers