# LLM Optimization

Jordan Boyd-Graber

University of Maryland

Quantization

Slides adapted from `https://huggingface.co/docs/optimum/concept_guides/quantization`

# Motivation for Quantization

### Why Quantize?

- **Memory efficiency:** reduces model size
- **Faster inference:** accelerates computation on specialized hardware
- **Lower energy use:** fewer bits mean less data movement
- **Deployment ease:** run on edge or CPU devices

### Why Not Quantize?

- **Accuracy loss:** rounding and clipping errors
- **Sensitivity:** some layers (e.g., embeddings, layernorm) don't work
- **Hardware dependence:** performance gains vary by platform
- **Complexity:** requires calibration or quantization-aware retraining

# LLAMA Example (from Meta)

- For models trained in FP16 (16-bit), converting to INT8 (8-bit) reduces memory usage by 50%, while INT4 (4-bit) reduces it by 75%.
- INT8 quantization can provide a 2-4x speedup on modern hardware, while INT4 can offer even greater speedups.

# Types of Quantization: `float16` vs. `int8`

## Reduced-Precision Floating Point (FP16)

$$x = (-1)^s \times (1 + m) \times 2^{e-15}$$

| Component | Bits | Range | Notes |
|---|---|---|---|
| Sign | 1 | $\pm$ | Sign bit |
| Exponent | 5 | $[-14, +15]$ | Reduced range |
| Mantissa | 10 | — | Reduced precision |

## Integer Quantization (INT8)

| Component | Bits | Notes |
|---|---|---|
| Sign + Magnitude | 8 | Two's complement integer |

# Types of Quantization: `float16` vs. `int8`

## Reduced-Precision Floating Point (FP16)

$$x = (-1)^s \times (1+m) \times 2^{e-15}$$

| Component | Bits | Range | Notes |
|---|---|---|---|
| Sign | 1 | $\pm$ | Sign bit |
| Exponent | 5 | $[-14, +15]$ | Reduced range |
| Mantissa | 10 | — | Reduced precision |

## Integer Quantization (INT8)

| Component | Bits | Notes |
|---|---|---|
| Sign + Magnitude | 8 | Two's complement integer |

# Affine Quantization: Mapping Float Values to INT8

**Core Equation**

$$x = S \times \left(x_q - Z\right)$$

- $x$ — original **floating-point** value.

$$x_q \in [-128, 127], \quad S > 0, \quad Z \in \mathbb{Z}$$

# Affine Quantization: Mapping Float Values to INT8

**Core Equation**

$$x = S \times \left(x_q - Z\right)$$

- $x$ — original **floating-point** value.
- $S$ — **scale factor** that maps integer steps to real value intervals.

$$x_q \in [-128, 127], \quad S > 0, \quad Z \in \mathbb{Z}$$

# Affine Quantization: Mapping Float Values to INT8

**Core Equation**

$$x = S \times \left(x_q - Z\right)$$

- $x$ — original **floating-point** value.
- $S$ — **scale factor** that maps integer steps to real value intervals.
- $x_q$ — **quantized integer** representation (usually in $[-128, 127]$).

$$x_q \in [-128, 127], \quad S > 0, \quad Z \in \mathbb{Z}$$

# Affine Quantization: Mapping Float Values to INT8

**Core Equation**

$$x = S \times \left(x_q - Z\right)$$

- $x$ — original **floating-point** value.
- $S$ — **scale factor** that maps integer steps to real value intervals.
- $x_q$ — **quantized integer** representation (usually in $[-128, 127]$).
- $Z$ — **zero-point offset**, ensuring $x = 0$ maps to an integer in range.

$$x_q \in [-128, 127], \quad S > 0, \quad Z \in \mathbb{Z}$$

# Affine Quantization: Mapping Float Values to INT8

**Core Equation**

$$x = S \times \left(x_q - Z\right)$$

- $x$ — original **floating-point** value.
- $S$ — **scale factor** that maps integer steps to real value intervals.
- $x_q$ — **quantized integer** representation (usually in $[-128, 127]$).
- $Z$ — **zero-point offset**, ensuring $x = 0$ maps to an integer in range.

**Typical ranges:**

$$x_q \in [-128, 127], \quad S > 0, \quad Z \in \mathbb{Z}$$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

## Setup

- Float range: $x_{\min} = -1.0$, $x_{\max} = 2.0$
- Integer range: $q_{\min} = -128$, $q_{\max} = 127$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

## Setup

- Float range: $x_{\min} = -1.0$, $x_{\max} = 2.0$
- Integer range: $q_{\min} = -128$, $q_{\max} = 127$
- Compute scale:

$$S = \frac{x_{\max} - x_{\min}}{q_{\max} - q_{\min}} = \frac{3}{255} \approx 0.01176$$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

## Setup

- Float range: $x_{\min} = -1.0$, $x_{\max} = 2.0$
- Integer range: $q_{\min} = -128$, $q_{\max} = 127$
- Compute scale:

$$S = \frac{x_{\max} - x_{\min}}{q_{\max} - q_{\min}} = \frac{3}{255} \approx 0.01176$$

- Compute zero-point:

$$Z = \mathrm{round}\left(q_{\min} - \frac{x_{\min}}{S}\right) = \mathrm{round}(-128 + 85) = -43$$

# Worked Example: Forward Quantization ($x \to x_q$)

Setup

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

# Worked Example: Forward Quantization ($x \rightarrow x_q$)

Setup

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$x = -0.8$$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$x = -0.8 \quad \longrightarrow$$

## Worked Example: Forward Quantization ($x \to x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$x = -0.8 \quad \to \quad \text{round}\left(\tfrac{-0.8}{0.01176} - 43\right)$$

# Worked Example: Forward Quantization ($x \rightarrow x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \mathrm{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$x = -0.8 \quad \rightarrow \quad \mathrm{round}\left(\tfrac{-0.8}{0.01176} - 43\right) \quad = -111$$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$x = -0.8 \quad \longrightarrow \quad \text{round}\left(\tfrac{-0.8}{0.01176} - 43\right) \quad = -111$$
$$x = 0.5$$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

Setup

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$\begin{aligned} x &= -0.8 \quad \longrightarrow \quad \text{round}\left(\tfrac{-0.8}{0.01176} - 43\right) \quad = -111 \\ x &= 0.5 \quad\;\; \longrightarrow \end{aligned}$$

# Worked Example: Forward Quantization ($x \rightarrow x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$\begin{aligned} x = -0.8 &\quad \rightarrow \quad \text{round}\left(\tfrac{-0.8}{0.01176} - 43\right) \quad = -111 \\ x = 0.5 &\quad \rightarrow \quad \text{round}\left(\tfrac{0.5}{0.01176} - 43\right) \end{aligned}$$

# Worked Example: Forward Quantization ($x \to x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$\begin{aligned}
x = -0.8 \quad &\to \quad \text{round}\left(\tfrac{-0.8}{0.01176} - 43\right) \quad &= -111 \\
x = 0.5 \quad &\to \quad \text{round}\left(\tfrac{0.5}{0.01176} - 43\right) \quad &= -1
\end{aligned}$$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$\begin{array}{llll} x = -0.8 & \longrightarrow & \text{round}\left(\frac{-0.8}{0.01176} - 43\right) & = -111 \\ x = 0.5 & \longrightarrow & \text{round}\left(\frac{0.5}{0.01176} - 43\right) & = -1 \\ x = 1.5 & & & \end{array}$$

# Worked Example: Forward Quantization ($x \to x_q$)

**Setup**

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$\begin{array}{lcll}
x = -0.8 & \longrightarrow & \text{round}\left(\frac{-0.8}{0.01176} - 43\right) & = -111 \\
x = 0.5 & \longrightarrow & \text{round}\left(\frac{0.5}{0.01176} - 43\right) & = -1 \\
x = 1.5 & \longrightarrow & &
\end{array}$$

# Worked Example: Forward Quantization ($x \longrightarrow x_q$)

Setup

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$\begin{aligned}
x &= -0.8 & \longrightarrow \quad & \text{round}\left(\tfrac{-0.8}{0.01176} - 43\right) & = -111 \\
x &= 0.5 & \longrightarrow \quad & \text{round}\left(\tfrac{0.5}{0.01176} - 43\right) & = -1 \\
x &= 1.5 & \longrightarrow \quad & \text{round}\left(\tfrac{1.5}{0.01176} - 43\right) &
\end{aligned}$$

# Worked Example: Forward Quantization ($x \to x_q$)

Setup

- $S = 0.01176$
- $Z = -43$

Forward mapping:

$$x_q = \text{round}\left(\frac{x}{S} + Z\right)$$

**Examples:**

$$\begin{aligned}
x &= -0.8 && \longrightarrow && \text{round}\left(\tfrac{-0.8}{0.01176} - 43\right) && = -111 \\
x &= 0.5 && \longrightarrow && \text{round}\left(\tfrac{0.5}{0.01176} - 43\right) && = -1 \\
x &= 1.5 && \longrightarrow && \text{round}\left(\tfrac{1.5}{0.01176} - 43\right) && = 85
\end{aligned}$$

# But how to set the mapping? (post hoc)

- Area of active research
- You have to quantize each layer
- Error propagates
- Strategy: Select the quantization that minimizes the overall error

# But how to set the mapping? (post hoc)

- Area of active research
- You have to quantize each layer
- Error propagates
- Strategy: Select the quantization that minimizes the overall error
  - On a representative data
  - Order matters!

# Recap

- Modern models are big
- Quantization lets you run them on smaller computers (or phones)
- While it degrades accuracy, you can strategically quantize to minimize it