Quantization

1. Full Precision / Half Precision → Data → Weights and Parameters
2. Calibration → Model Quantization → Problems
3. Modes of Quantization
   - ↳ Post Training Quantization
   - ↳ Quantization Aware Training

→ Quantization

Quantization is a compression technique that involves mapping high precision values to a lower precision ones.

2.52	-1.12	1.74	0.05
0.08	-0.22	-1.21	2.65
-0.13	1.60	0.02	-1.31
2.13	-0.01	1.83	1.65

32 bit float

→

121	-54	83	2
4	-11	-58	127
-6	77	1	-63
102	0	88	79

8 bit int

> Full Precision / Half Precision

$$\begin{tikzcd}
FP32 \arrow[r, "Quantization"] & FP16 \\
\text{Full Precision} \arrow[u] & \text{Half Precision} \arrow[u]
\end{tikzcd}$$

> ~~Calibration~~ How to Perform Quantization

— Symmetric Quantization

└> Batch Normalization

[0.0 .... 1000] ——> Numbers ——> 32 bits

└ 0-255 => 8 bits

Min Max Scaler

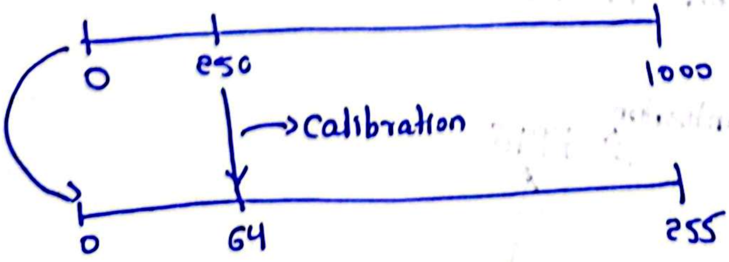
0.0 ——> 0

1000 ——> 255

$$\text{Scale} = \frac{x_{max} - x_{min}}{q_{max} - q_{min}}$$

$$= \frac{1000 - 0}{255 - 0}$$

$$= 3.92$$



$$\text{round}\left(\frac{250}{3.92}\right) = 64$$

– Asymmetric Quantization

$$[-20.0 \text{ ------ } 1000]$$

$$[0 \text{ ------ } 255]$$

$$\frac{1000 \mp (-20)}{255 - 0} = \frac{1000+20}{255} = 4.0 \Rightarrow \text{Scale factor}$$

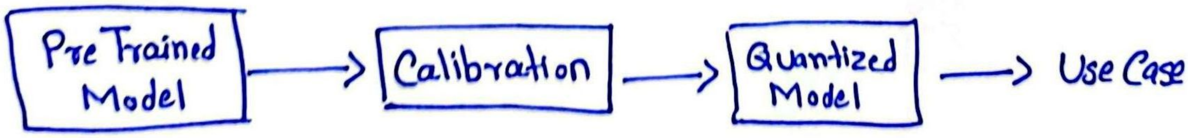
$$\text{round}\left(\frac{-20}{4}\right) = -5.0$$

$$\Downarrow$$

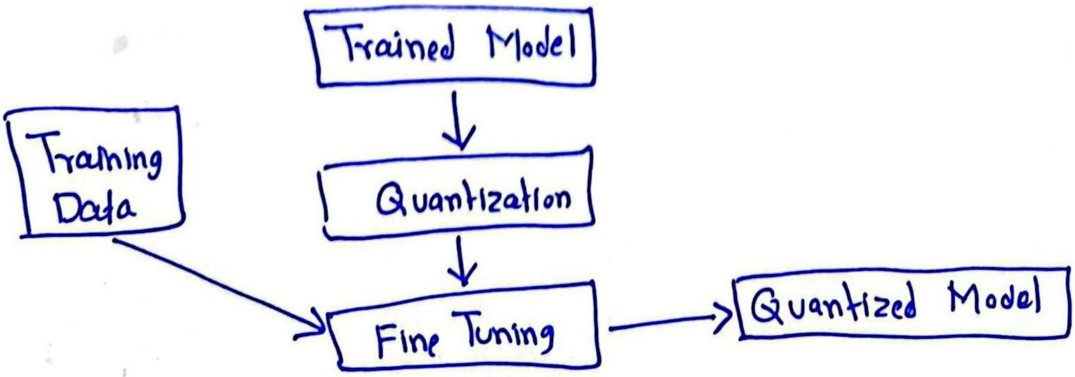
$$-5.0 + \underbrace{5}_{\text{zero point}} = 0$$

> Modes of Quantization

– Post Training Quantization



– Quantization Aware Training (QAT)


Trained Model
Training Data
Quantization
Fine Tuning
Quantized Model