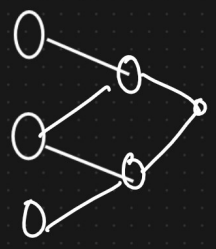
Fine Tuning LLM Models

QLORA / LORA

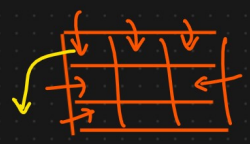
① Quantization

① Full Precision / Half Precision → DATA → Weights and parameter ✓
② Calibration — Model Quantization → Problems ✓
③ Modes of Quantization → Post Training Quantization ✓
→ Quantization Aware Training ✓

Quantization : Conversion from higher memory format to a lower memory format.



Weights → Matrix



32 bits

FP 32 → Full Precision / Single Precision.

floating points 32 bits

7.23 ⇒ 32 bits ⇒ int 8 ⇒ Inference → Mobile phone ✓
→ Edge Device ✓
→ Watches Smart
⇒ Quantization → Fine Tuning → loss of Accuracy

LLama 2 70 billion

Cost ↑↑ → Resources

TF 32 bit

FP 32 bit ⇒ FP 16 bit
↳ Half Precision

② How to perform Quantization

① Symmetric Quantization ② Asymmetric Quantization

↓ ↑

Batch Normalization

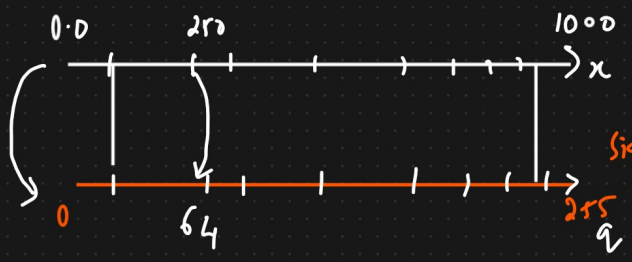
① Symmetric uint8 quantization

[0.0 ..... 1000] → Numbers → 32 bits → uint8 ⇒ 8 bits ⇒ 2^8

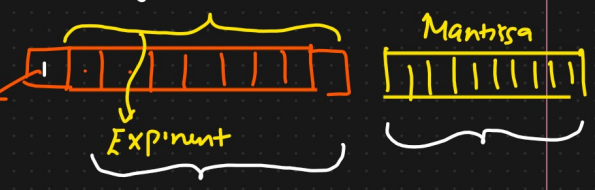
→ 0 - 255

0 - 255

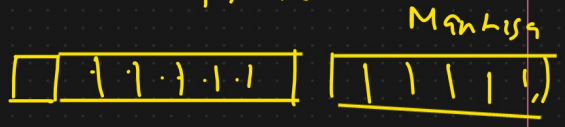
7.32



Single Precision Floating Point 32



FP 16



Min max Scalar

0.0 → 0

1000 → 255

$$Scale = \frac{x_{max} - x_{min}}{q_{max} - q_{min}} = \frac{1000 - 0}{255 - 0} = 3.92$$

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

Zero point = 0

② Asymmetric uint8 quantization

$$\frac{1000 + 20}{255} = 4.0 \Rightarrow \text{Scale factor}$$

[-20.0 - - - - - 1000.0]

-5.0

[0 - - - - - 255]

Zero point = 5

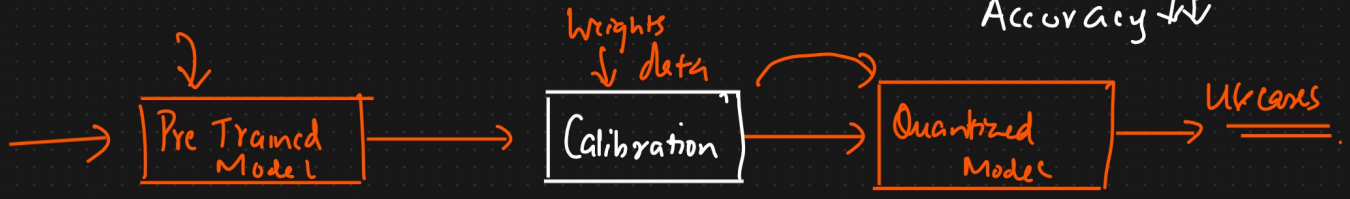
Scale = 4.0

$$round\left(\frac{-20}{4}\right) = (-5.0) + 5 = 0$$

-128 | 127

① Post Training Quantization (PTQ)

Loss of Data ⇓ Accuracy ↓



② Quantization Aware Training (QAT)