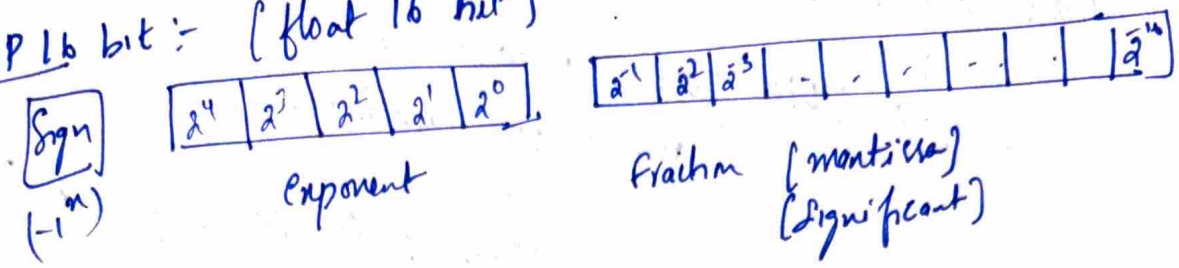
15/Oct/24

How does Quantization Work

https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-quantization#%C2%A7the-era-of-bit-llms-bitnet

→ IEEE 754 Standard of representing numbers

→ FP16 bit :- (float 16 bit)



$n = (0, 1)$

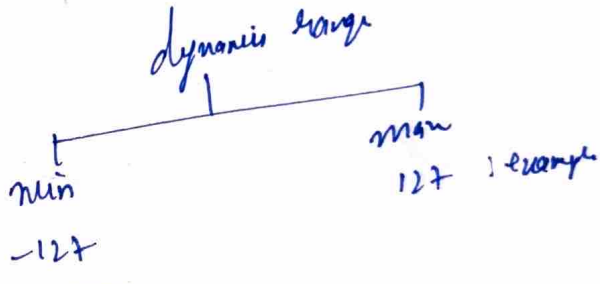
→ higher the bit value, higher the precision

→ Dynamic Range :- Interval of representable numbers a given representation

→ Precision :- Distance between two neighboring values

→ low distance = high precision

Inversely Proportional.



$$\text{memory} = \frac{\text{nr\_bits}}{8} \times \text{nr\_params.}$$

→ most models are in FP32 → Full Precision

70B Param model need ~~250~~ 280 GB of memory

70B model

32 Bit ⇒ $\frac{32}{8} \times 70B = 280$ GB

64 Bit ⇒ $\frac{64}{8} \times 70B = 560$ GB

16 Bit ⇒ $\frac{16}{8} \times 70B = 140$ GB

→ If Precision ↓ Accuracy of model ↓

→ but we should, focus on reducing bits while maintaining accuracy

→ int 8 → [-127 to 127]
  → 8 bit

→ FP32 → [$-3.4e^{38}$ to $3.4e^{38}$]
  → 32 bits.

Symmetric quantization =
symmetric around the zero.
Asymmetric quant = Vice versa

→ Absolute Maximum Quantization :- (absmax) form of symmetric Quantization

→ Scaling factor ⇒ $$S = \frac{2^{b-1} - 1}{\alpha}$$

b = # of bytes that we want to quantize to

α = highest absolute value

$$x_{quantized} = round[S \cdot x]$$

$$y = w \cdot x + b$$

w, b → static values

x → dynamic values

Activations [continiously gets updated after each hidden layer]

Calibrating Quantization

Methods :-

① (PTQ) Post Training Quantization :-
Quantization after training

② (QAT) Quantization Aware Training :-
" during training / fine tuning.

PTQ :- done after training Via Symmetric or Asymmetric quantization.

For Activations (x, y) → As we dont know their values or it requires inference of the model.

It requires ① Dynamic Quantization } for Activations (x, y)
② Static Quantization