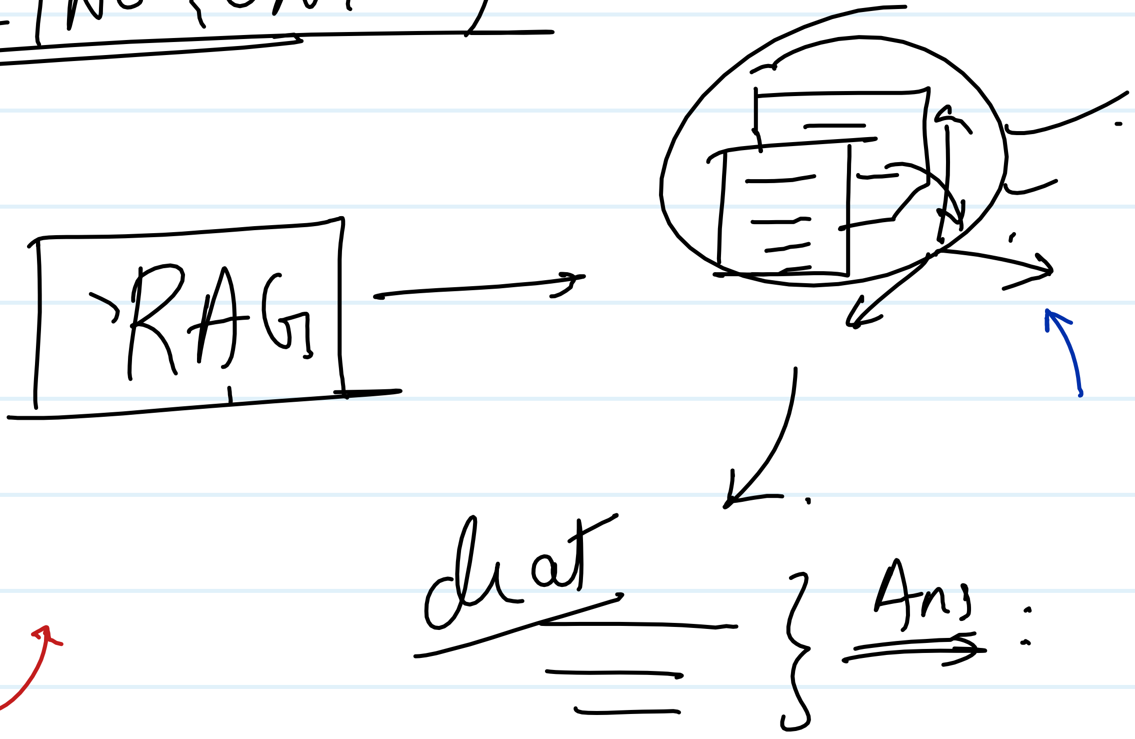
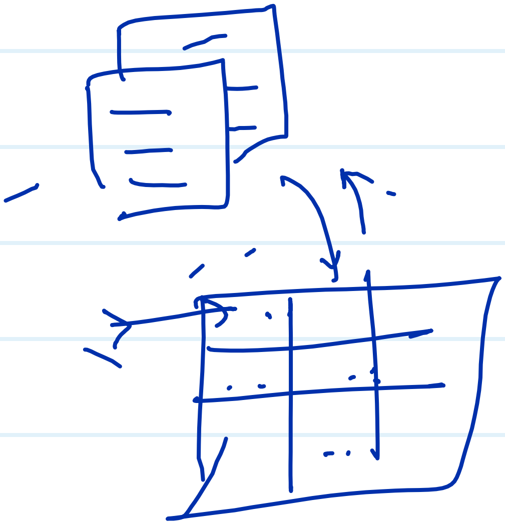
① → FINETUNING



FINETUNING :

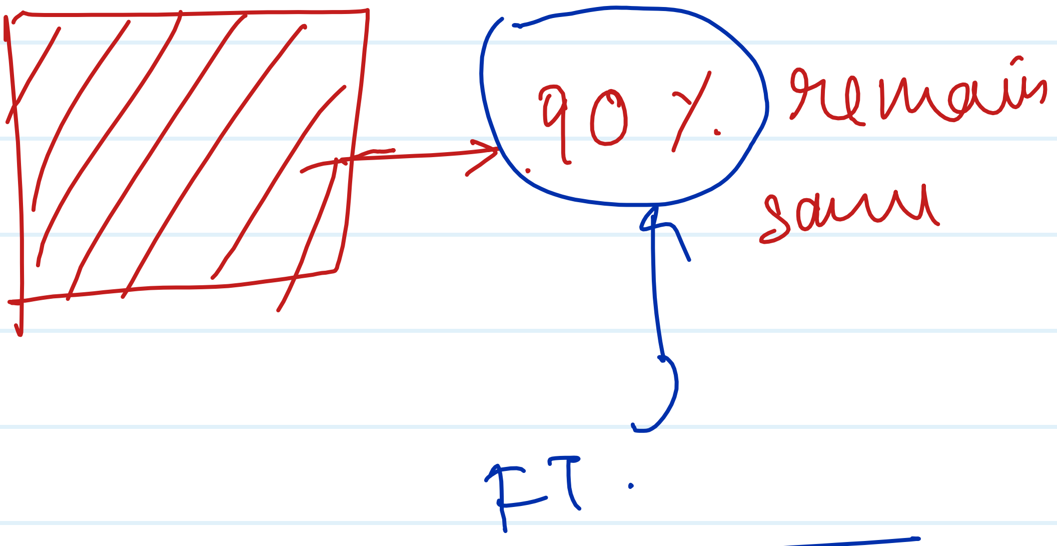


→ COMPLEX REASONING.

→ Specific output (JSON)

→ Interview

☆ PEFT ← {Parameter Efficient F.T..}.



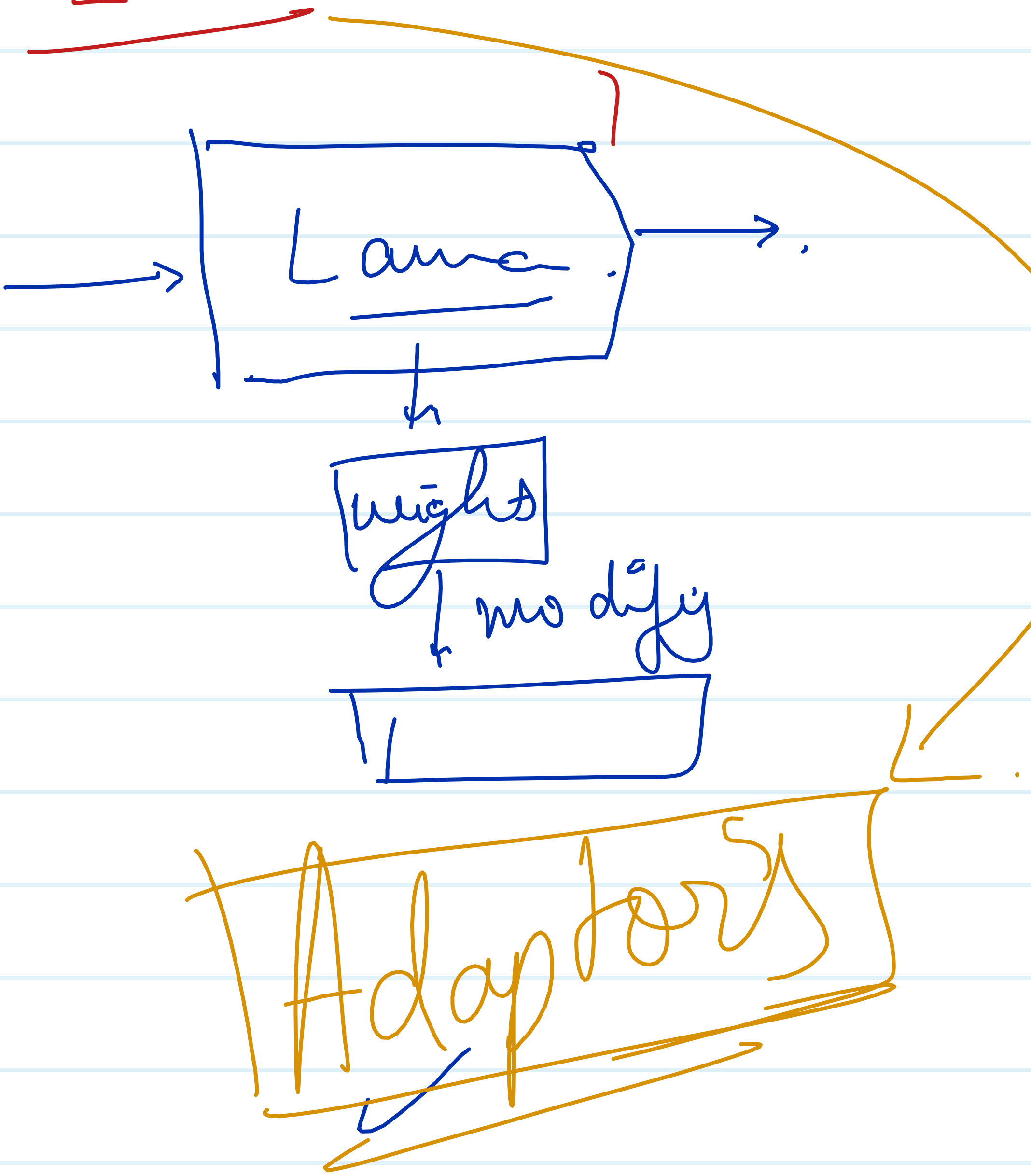
PEFT → LoRA

→ Low Rank Adapta^{n}.

→ QLoRA.

→ Quantized ———

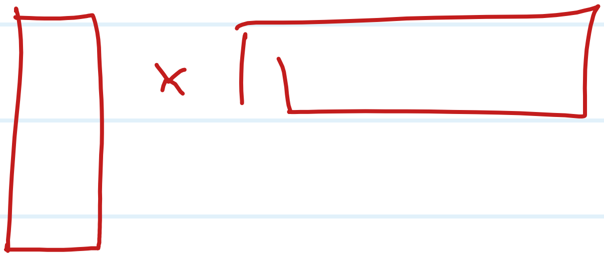
LoRA:



$$W_{new} = \underbrace{\underset{(n,m)}{W_{old}}} + \boxed{\Delta W}$$

Adaptors → $\underbrace{A}_{\text{Noise}} \times \underbrace{B}_{\text{Zero}}$.

{Matrix decompositions}.



W_{old} => (4096×4096).

$A \times B$ => $A \to (4096 \times \boxed{r})$

$B \to (\boxed{r} \times 4096)$.

Rank => $\underbrace{r = 1}$.

$A \to (4096 \times 1)$

$B \to (1 \times 4096)$

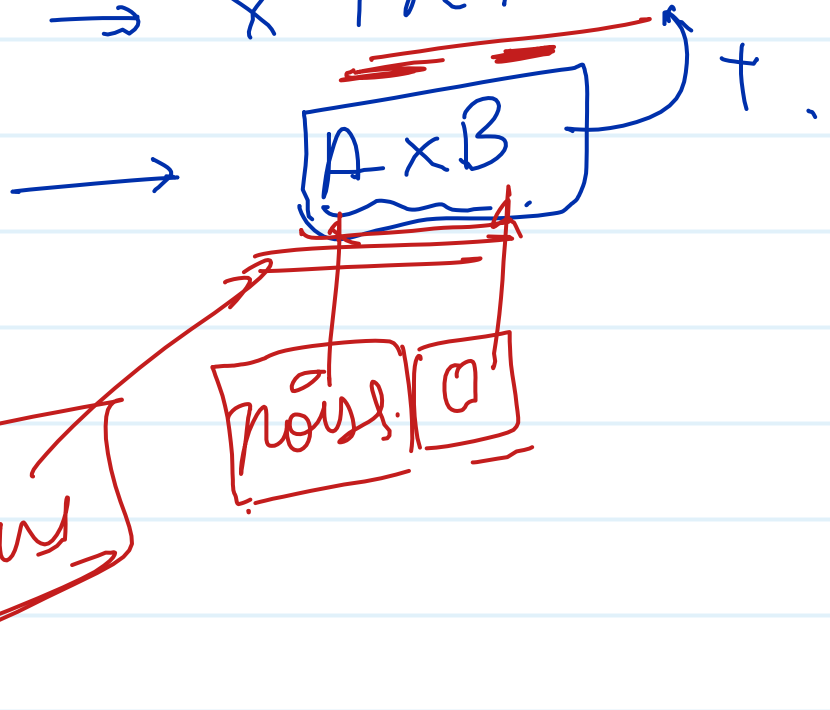
$$A \times B = 16 \text{ million}$$

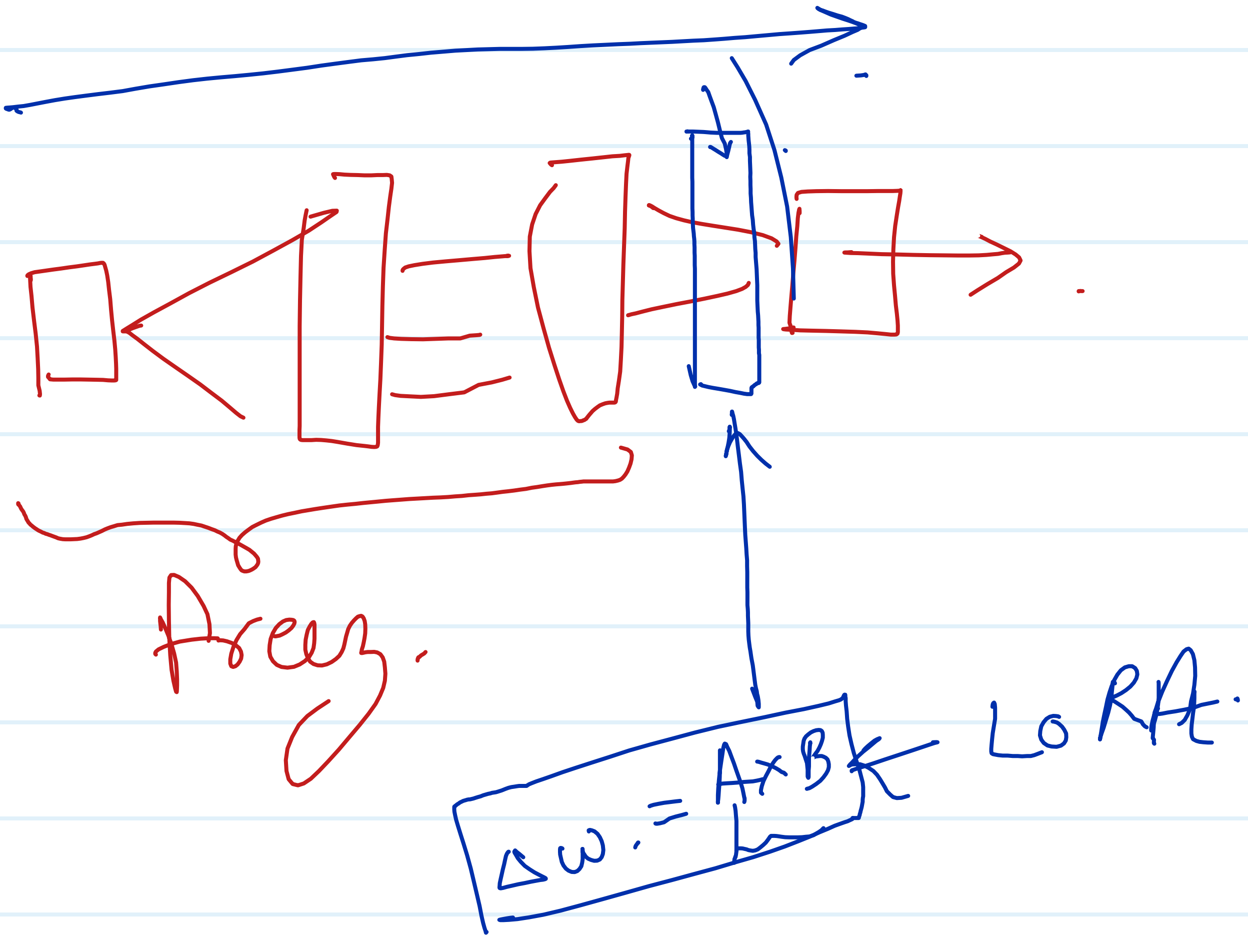
$(n \times \cancel{m})$ $(\cancel{m} \times p)$

1 1

A × B =>.

LoRA → × Pretrained (Frozen)





$A \rightarrow 4096 \times 1 = 4096$

$B \rightarrow 1 \times 4096 = + 4096.$

~ 9000

Pretrained → 4096×4096

⇒ =16million

Why is B = Zero M-1

zero ⇒ 0 → 0.001,

$$\underbrace{A \times B = 0}.$$

$$\underbrace{W_{new} = W_{old}} + \Delta W$$

$= 0$ ← no traing

Q+LoRA

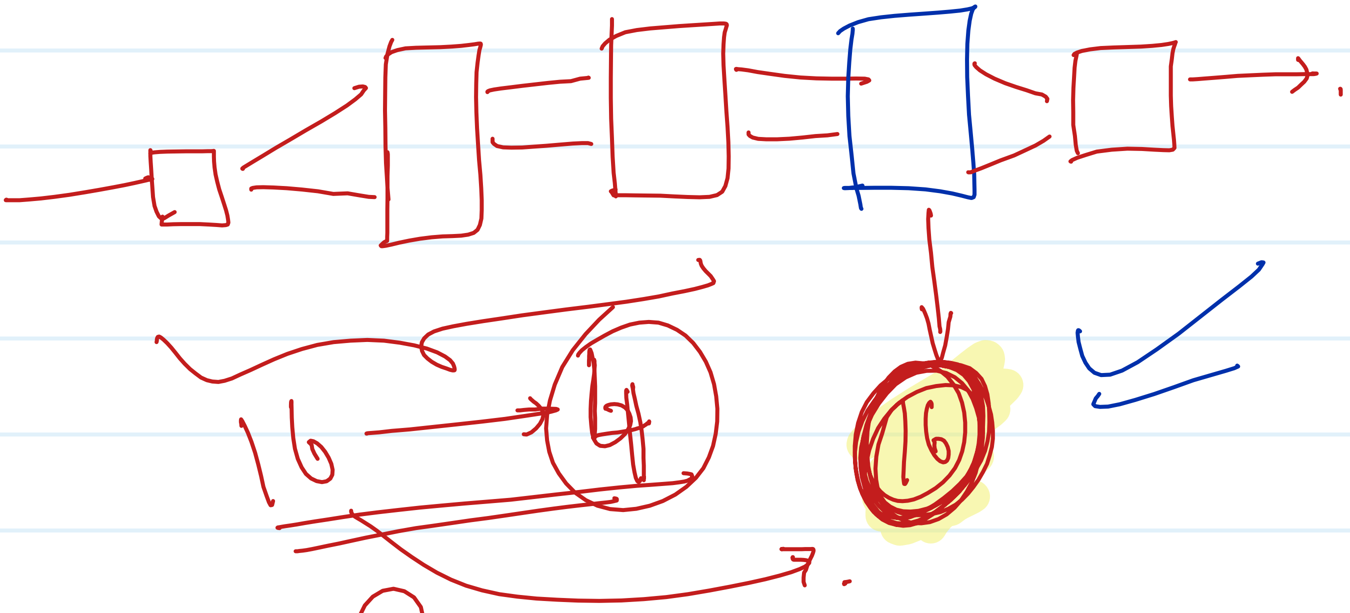
→ eq. → 16 bit

↓ Quantize

4bit

Quantize → Pretrained (eg. 16 → 4)

→ bits & bytes lib.



→ Storage

→ Computation..

chunks ← {10 0 4 ← pre ← fine
10 (16).

GPU ←

dequantize

4 → 16 ← instance

Final ⟹ $$w_n = \underbrace{w_0}_{} + \underset{\overline{F}}{\boxed{\frac{\boxed{\alpha}}{r}}}\,(\Delta w)$$

$$\uparrow$$

$$\underbrace{(A \times B)}_{}.$$

Thumb rule : $\alpha = 2r$.

$r \approx 16$.