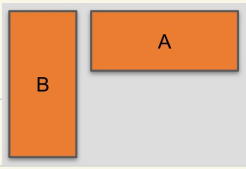
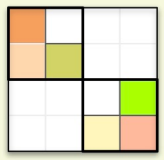


Methods :-

① LoRA $\quad \delta W = B \times A \quad B \in \mathbb{R}^{m \times r} \quad A \in \mathbb{R}^{r \times n}$

② SVD+LoRA $\quad \delta W = U \Sigma V^T \quad U \in \mathbb{R}^{m \times m} \quad \Sigma \in \mathbb{R}^{r \times r} \quad V \in \mathbb{R}^{n \times n}$

Sparse Fine Tuning :-



③ Fourier $\quad \delta W = F C F^T$

④ Hadamard $\quad \delta W = H C H^T$

⑤ Frame $\quad \delta W = F_r C F_r^T$

⑥ Random $\quad \delta W = R C R^T$

(F) = Inv. Fourier Transform

(H) = Inv. Hadamard Transform

(Fr) = Inv. frame Transform

(R) = Random values

* Computation :-

(*) We will be computing

$$Y = \delta W X \qquad X \in \mathbb{R}^{n \times d} \quad \text{(no. of tokens} \times \text{hidden dim)}$$

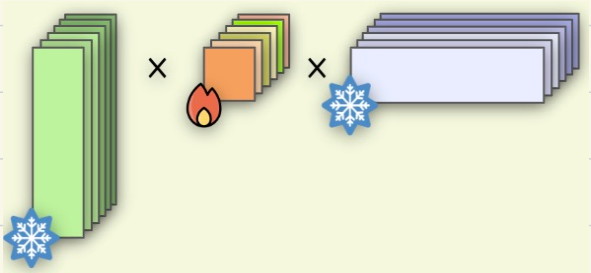
① Direct form : $\delta W X$

②

$$\sum_{i=1}^{k} P_{m,i} \; C_i \; P_{n,i}^T \; X$$

$P_{mi} \in \mathbb{R}^{m \times 1} \quad C_i \in \mathbb{R}^{1 \times 1} \quad P_{ni} \in \mathbb{R}^{n \times 1}$

③



$$\sum_{i} \sum_{j} P_{m,(i,j)} \; C_{(i,j)} \; P_{n,(i,j)}^T X$$

$P_{m,(i,j)} \in \mathbb{R}^{m \times 1} \quad C_{(i,j)} \in \mathbb{R} \quad P_{n,(i,j)} \in \mathbb{R}^{n \times 1}$

Models :-

① Llama-2-7B — m=2048, n=2048

② Gemma-2-2B — m=2048 n=2048

m=2304 n=1024

③ Gemma-2-9B — m=3584 n=4096

m=3584 n=2048

④ Llama-3.1-8B — m=4096 n=4096

m=4096 n=1024

⑤ Deep Seek !

Number of Tokens :-

→ n = 1, 256, 1024, 2048, 4096, 8192

Sparse blocks :-

① Size of C → l=2, 4, 16, 64, 256

② Sparsity of l → 75%, 50%, 25%, 0%

Precision :-

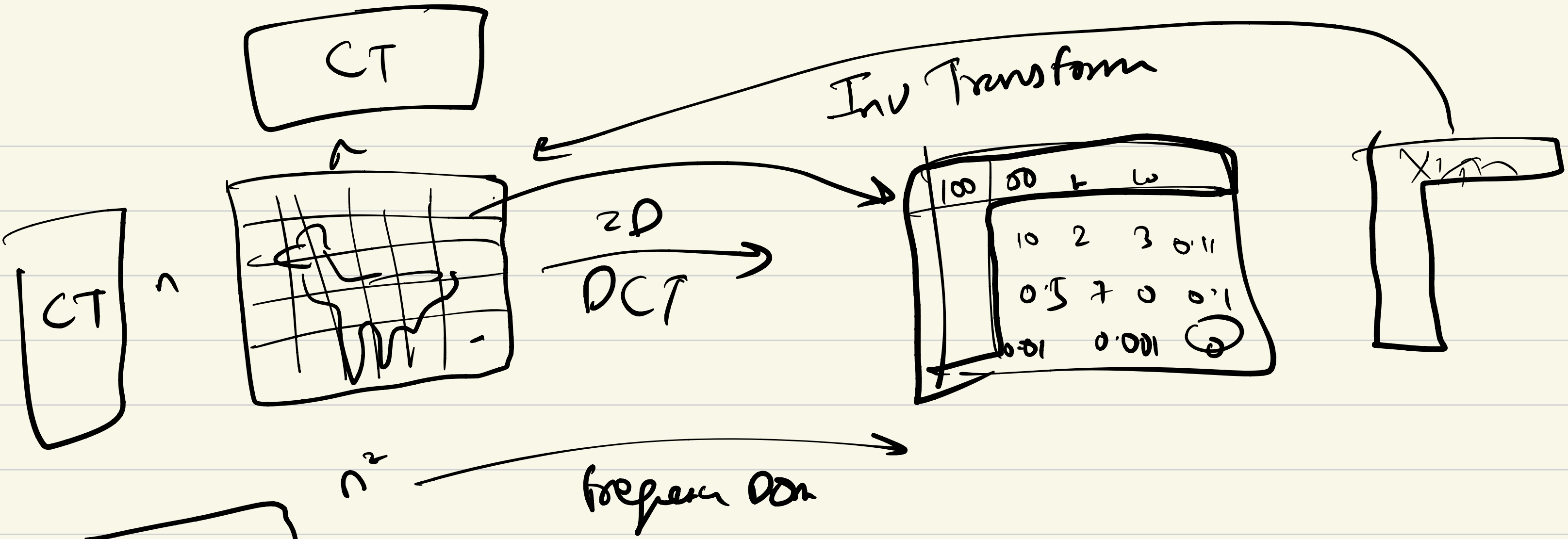
① FP32

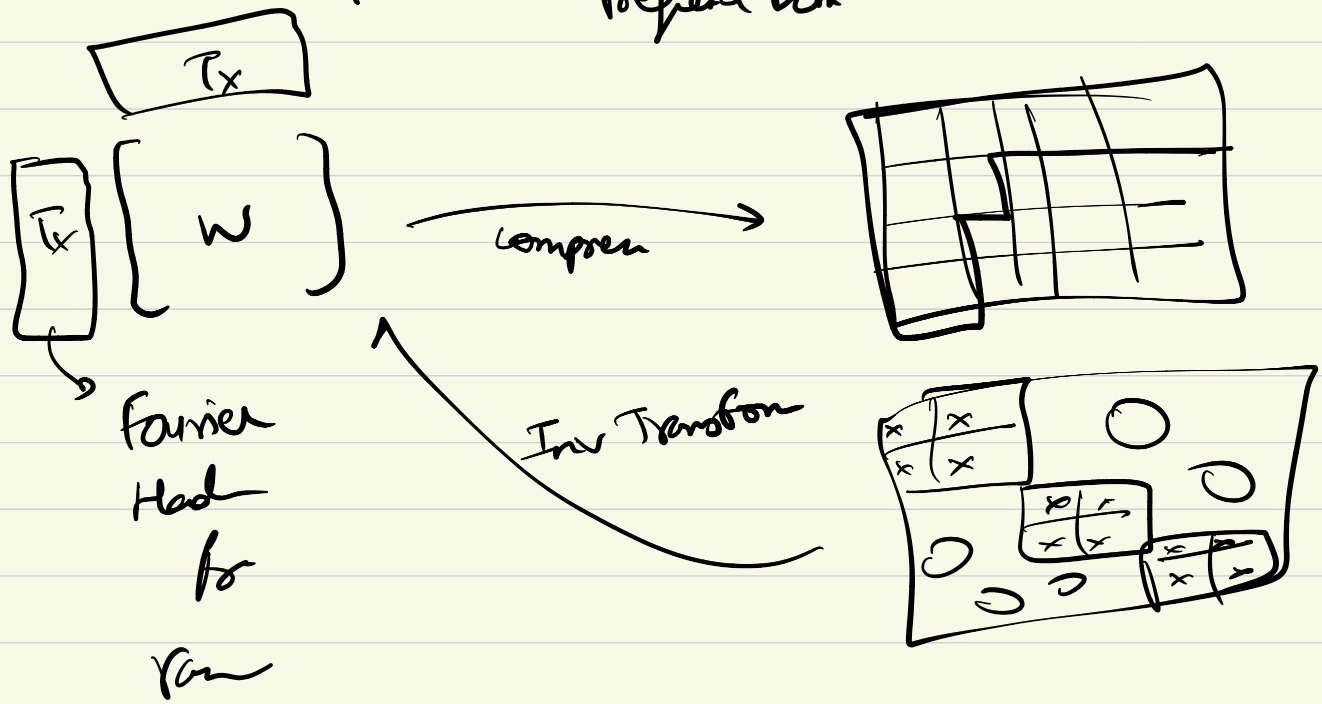
② FP16

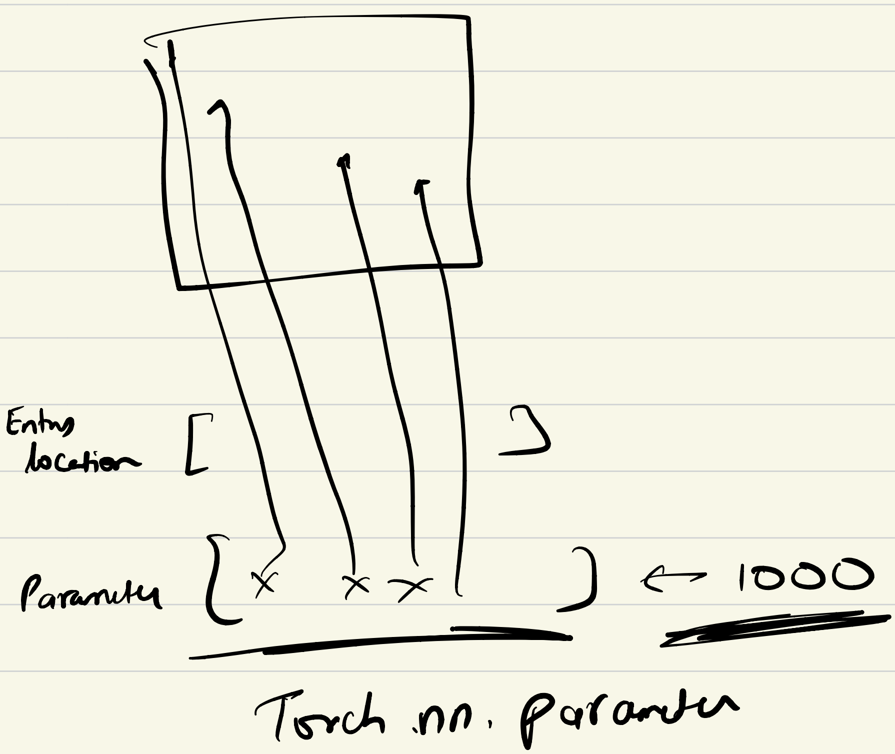
③ FP8

Threads :-

→ number of threads - 8, 16, 32, 256, 1024 etc.


CT
CT
2D
DCT
Inv Transform
100
10 2 3 0.11
0.5 7 0 0.1
0.01 0.001 0
n²
frequency Dom


Tx
Tx
W
compress
Inv Transform
Fourier
Head
for


Entry location
Parameter
1000
Torch.nn.parameter