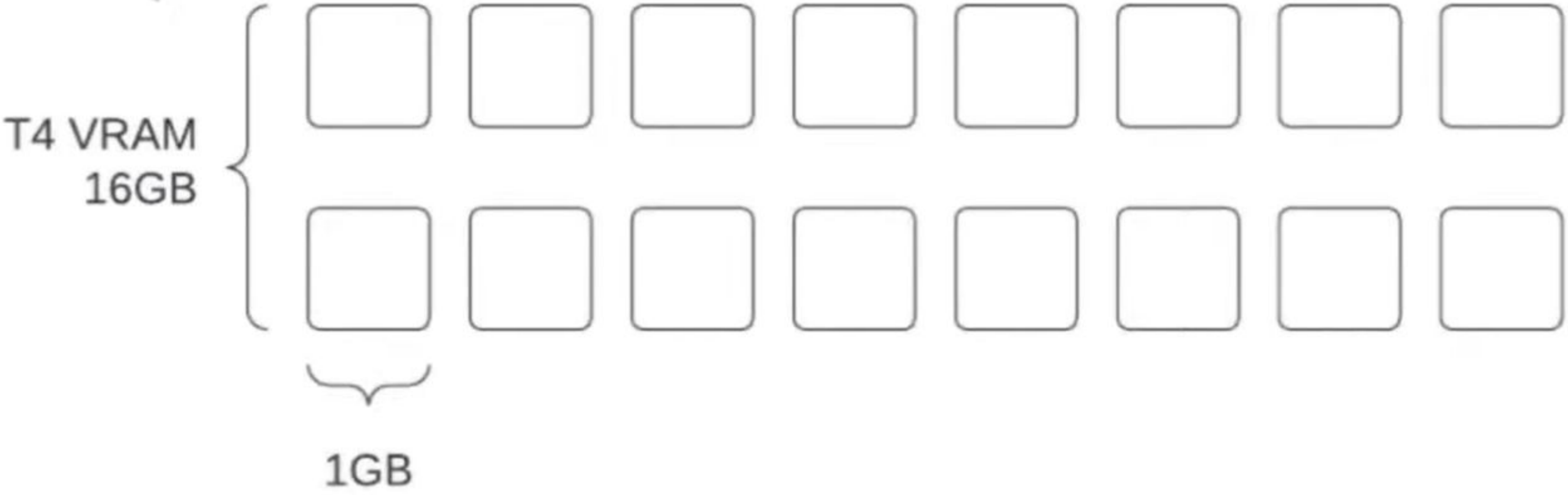
MEMORY OPTIMIZATION FOR FINE-TUNING MODELS

Sneha Jha
Imperial College London

ML in PL 2023

The VRAM Bottleneck

GPU	Tier	$ / hr (AWS)	VRAM (GiB)
H100	Enterprise	12.29	80
A100	Enterprise	5.12	80
V100	Enterprise	3.90	32
A10G	Enterprise	2.03	24
T4	Enterprise	0.98	16
RTX 4080	Consumer	N/A	16


T4 VRAM
16GB
1GB

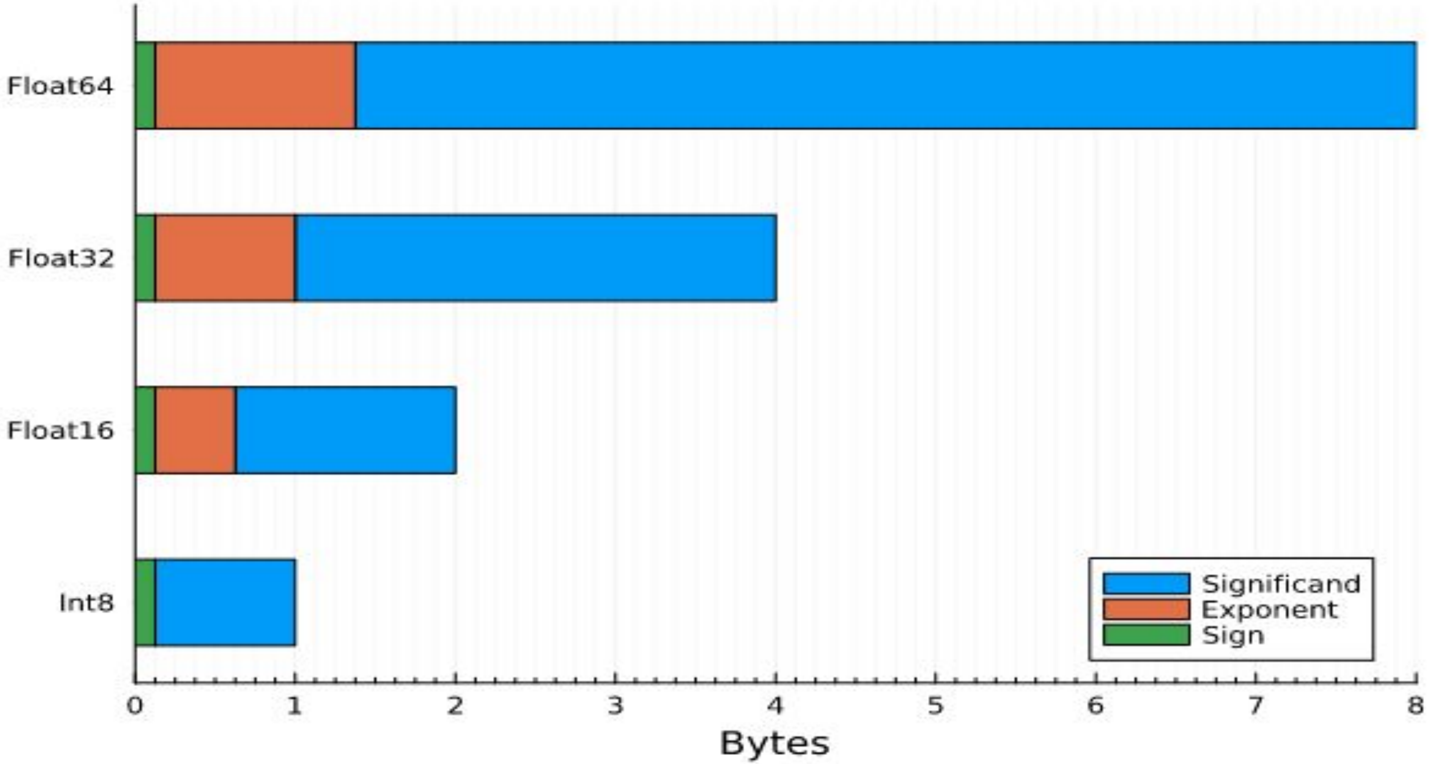
Sources of the Bottleneck

Model Parameters

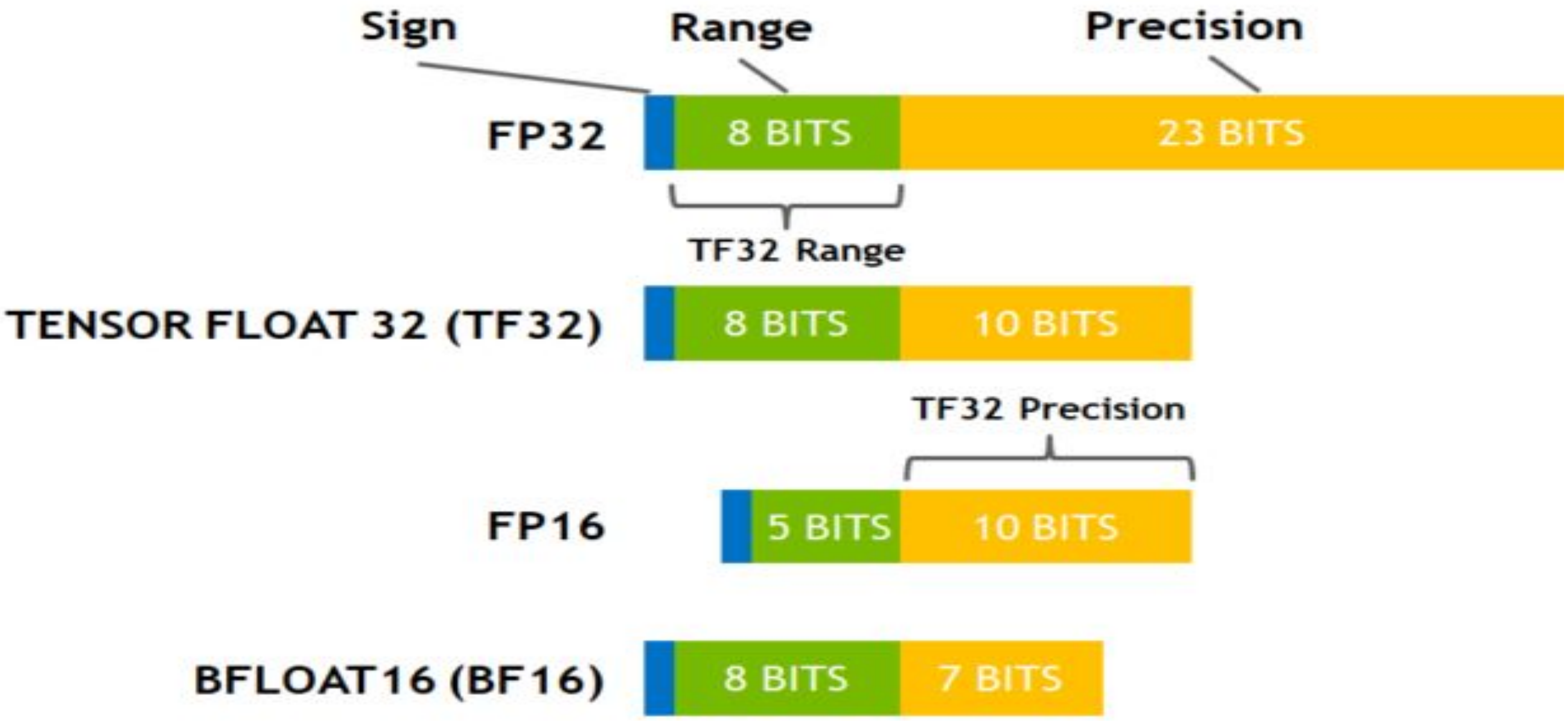
Gradients

Optimizer States

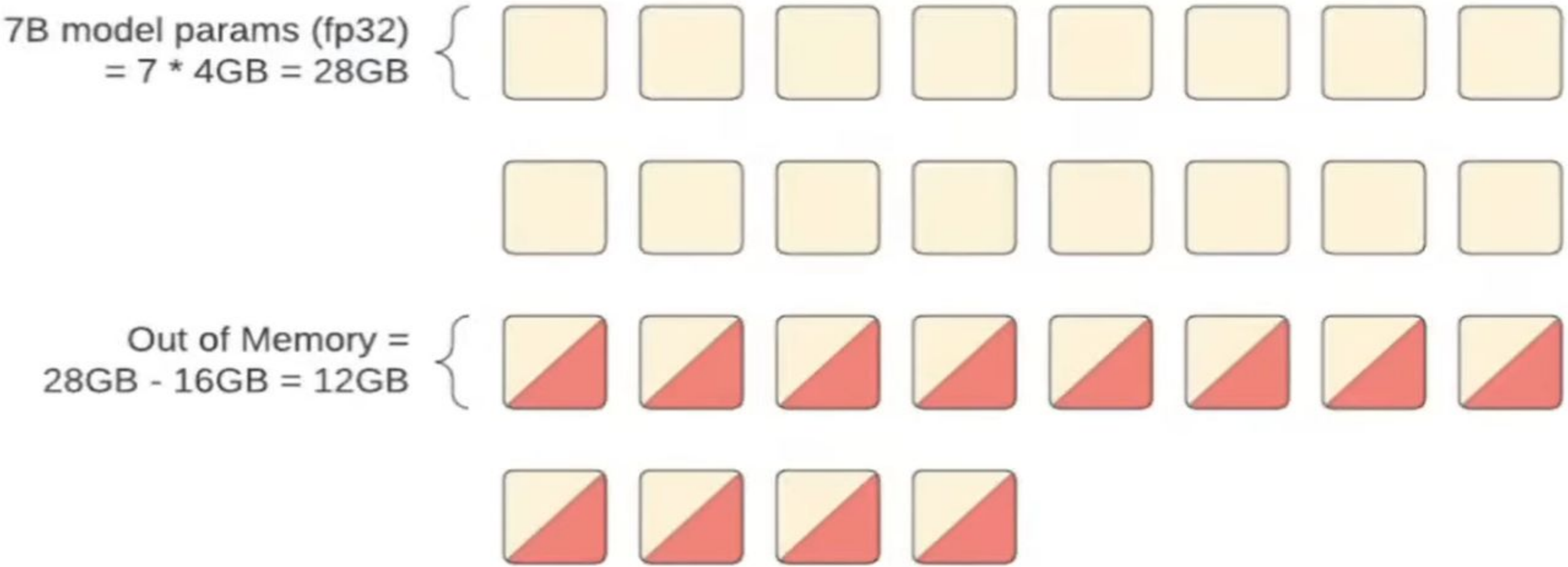
Precision Options



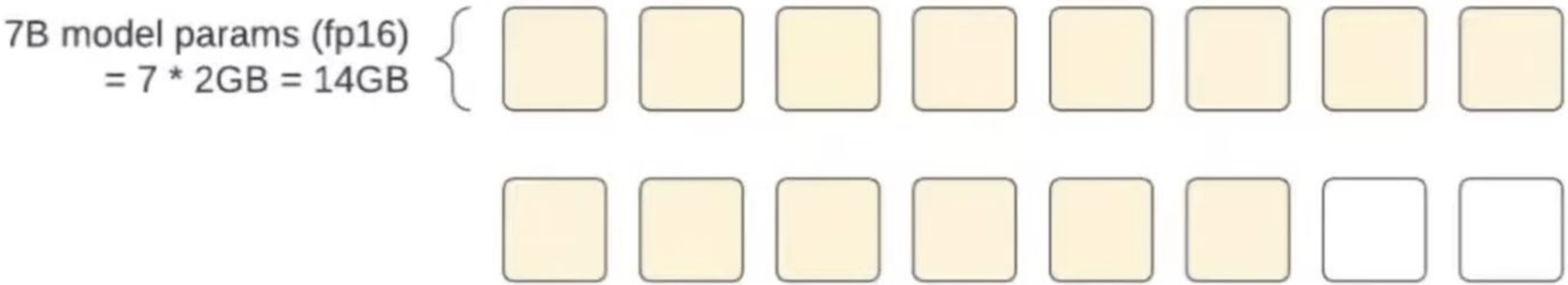
Precision Options



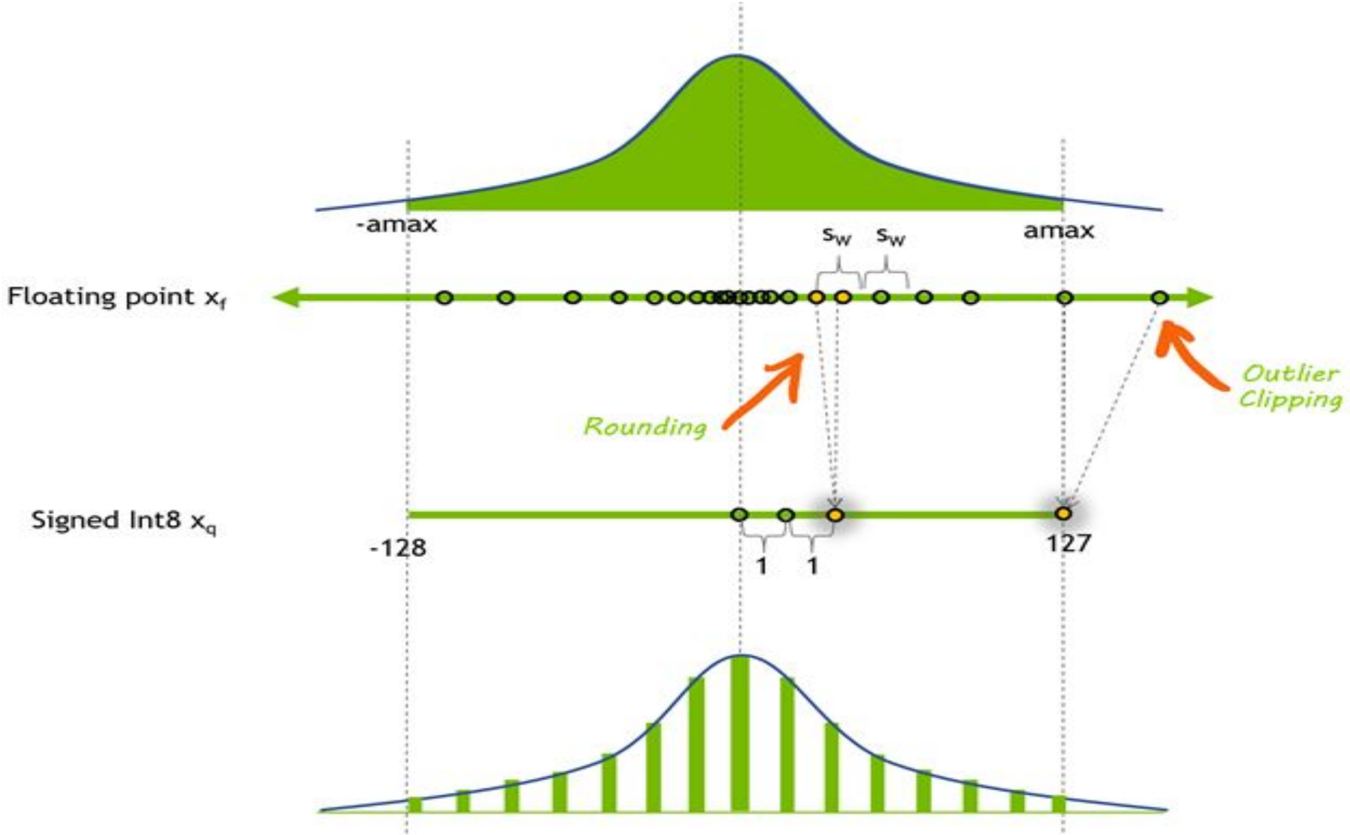
Model Parameters



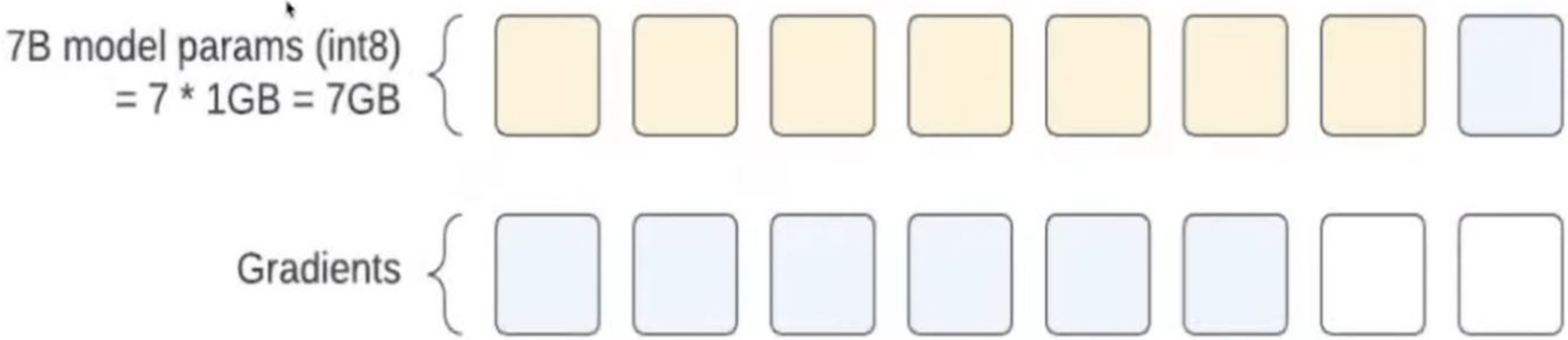
Model Parameters

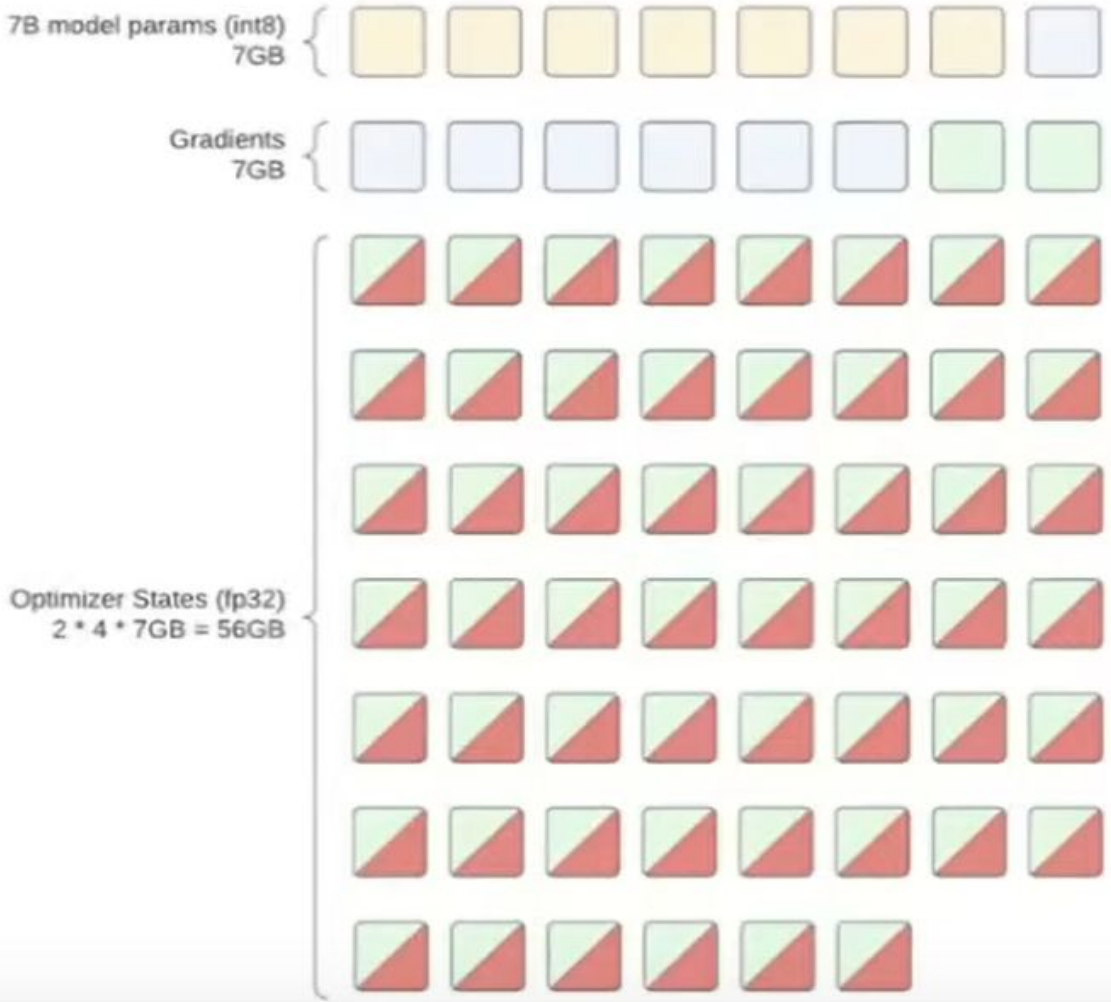


Quantization



Quantization




7B model params (int8)
7GB
Gradients
7GB
Optimizer States (fp32)
2 * 4 * 7GB = 56GB

Optimizer State

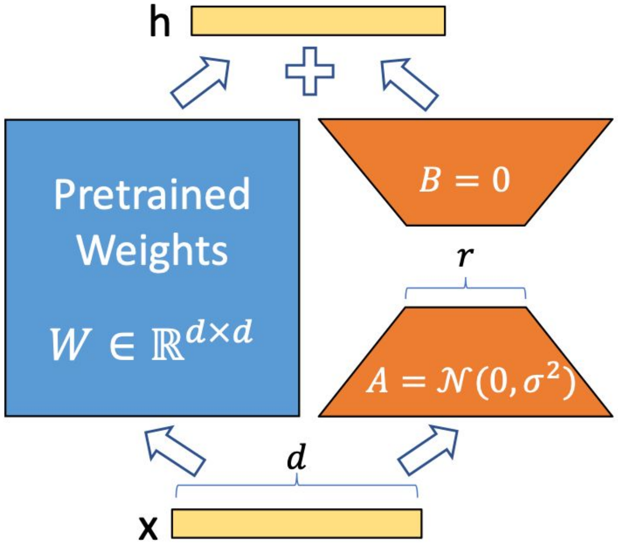
$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * \nabla w_t$$

$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * (\nabla w_t)^2$$

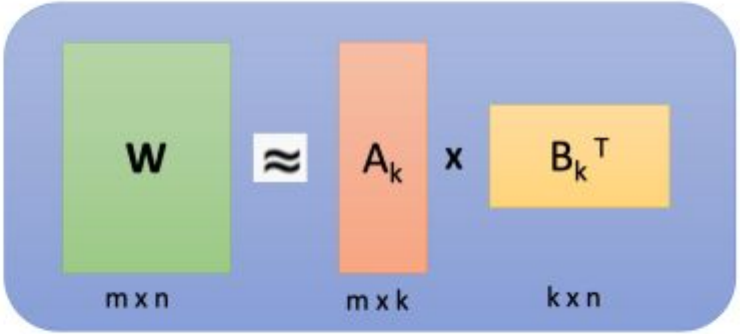
$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \qquad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} * \hat{m}_t$$

Low Rank Adaptation (LoRA)

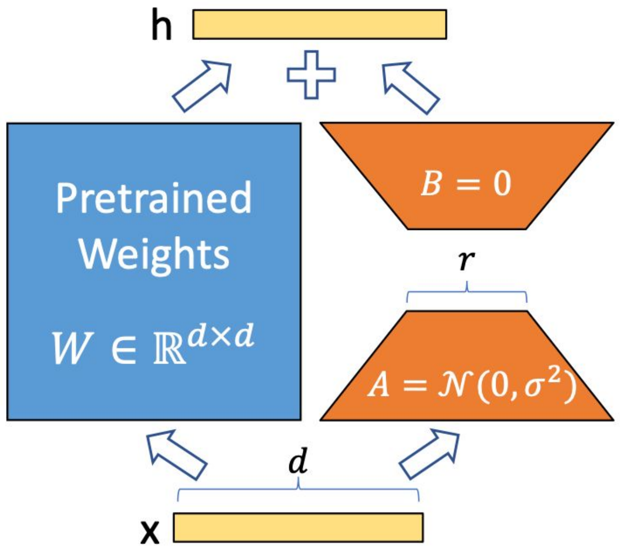


Low Rank Decomposition of a Matrix



Low Rank Adaptation (LoRA)

Weight matrix
1024 x 1024

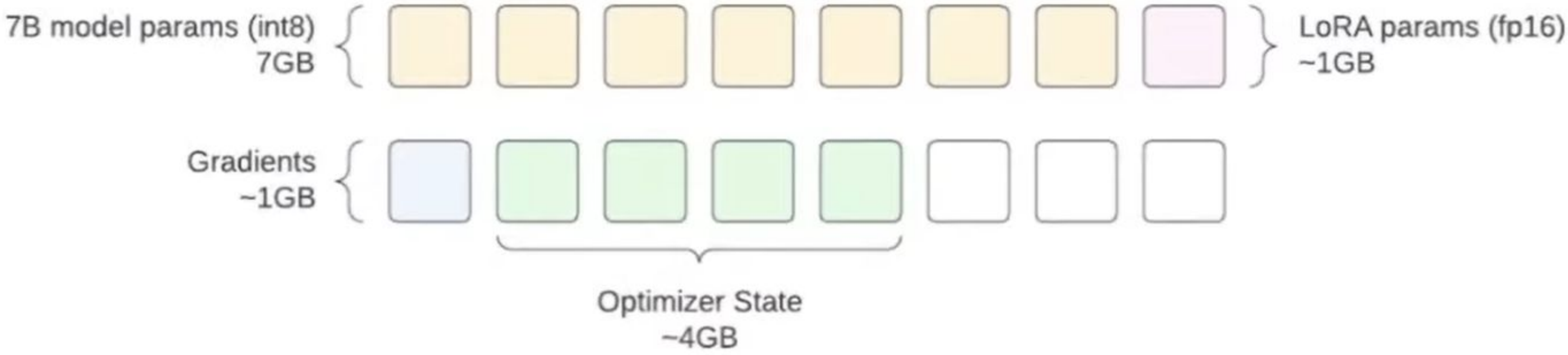


LoRA matrices
1024x8 and 8x1024
respectively

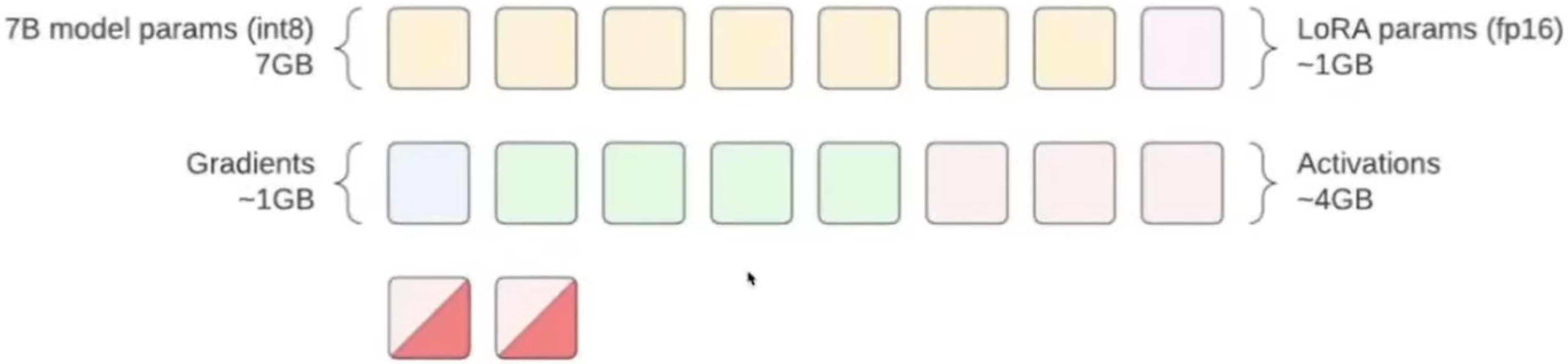
Uses 1024 * 8 * 2
parameters

Compresses tunable parameters

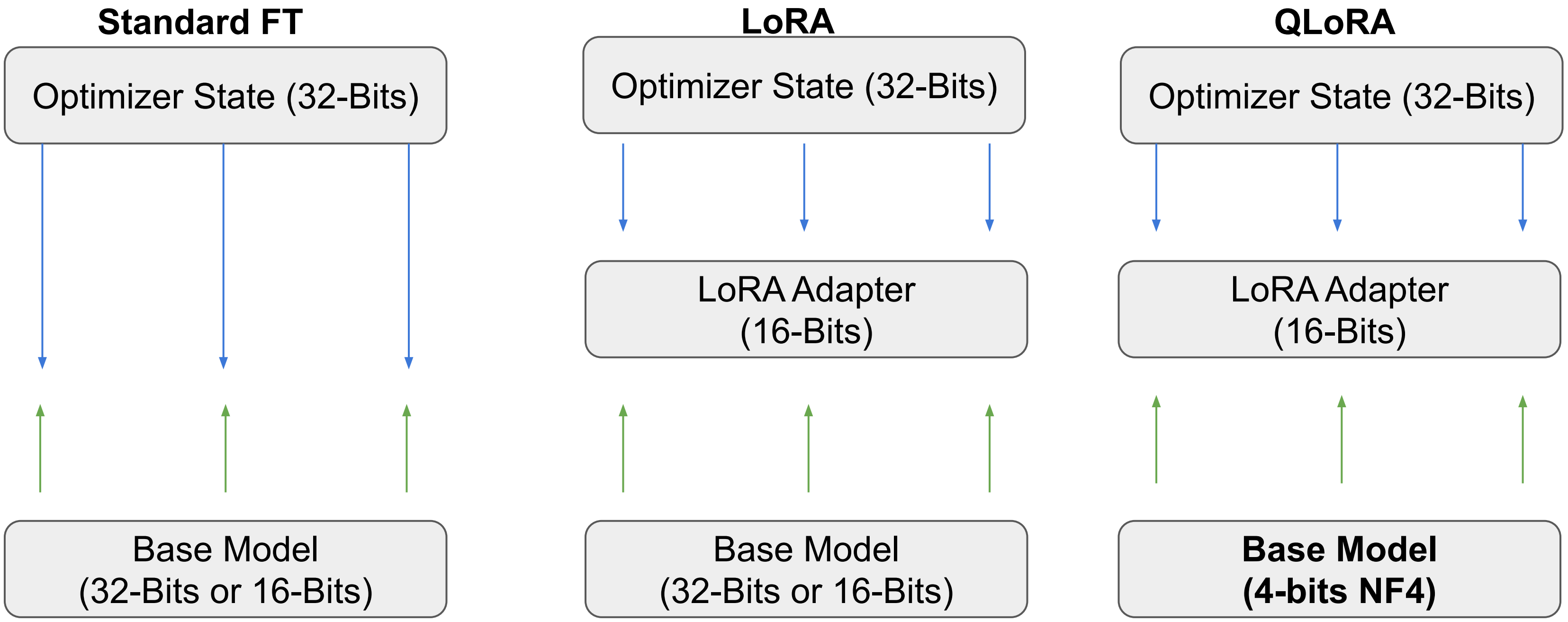
Low Rank Adaptation (LoRA)



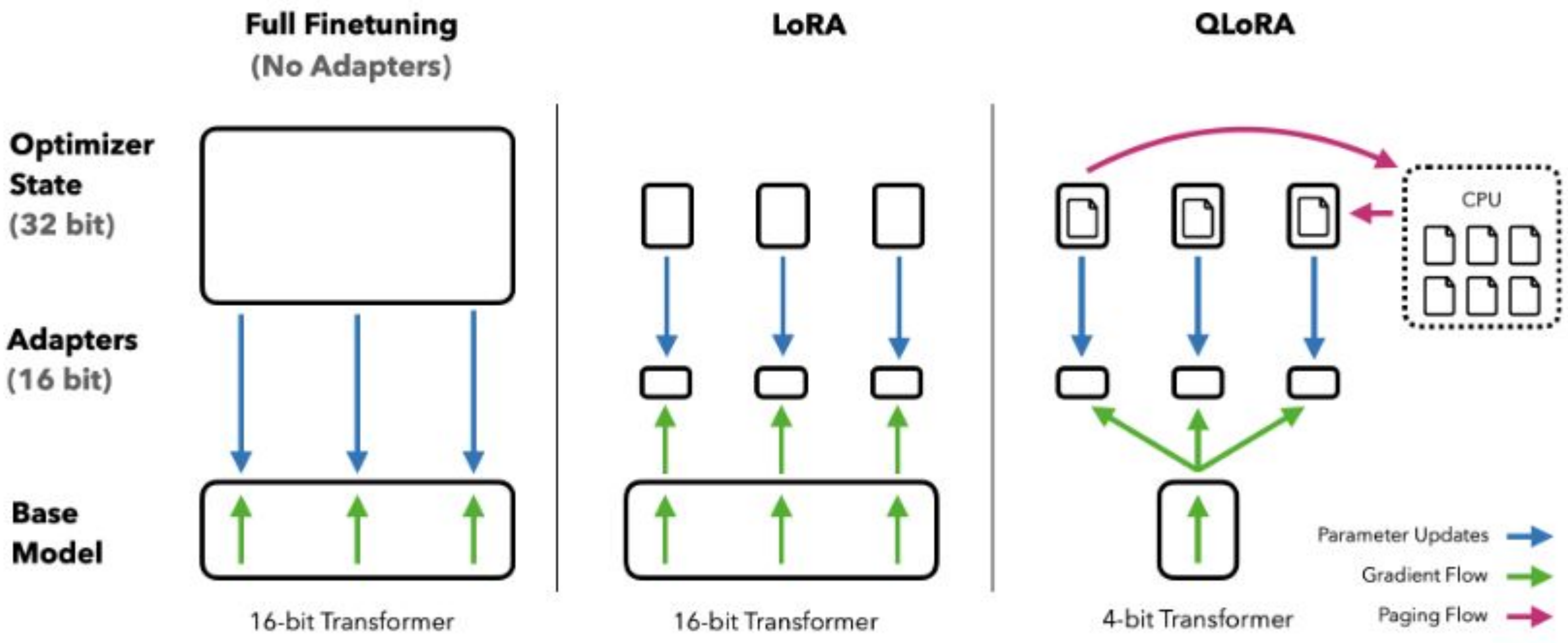
Activations



Quantized LoRA



QLoRA



Full Finetuning
(No Adapters)
LoRA
QLoRA
Optimizer
State
(32 bit)
Adapters
(16 bit)
Base
Model
CPU
16-bit Transformer
16-bit Transformer
4-bit Transformer
Parameter Updates
Gradient Flow
Paging Flow

Towards NF4 representation

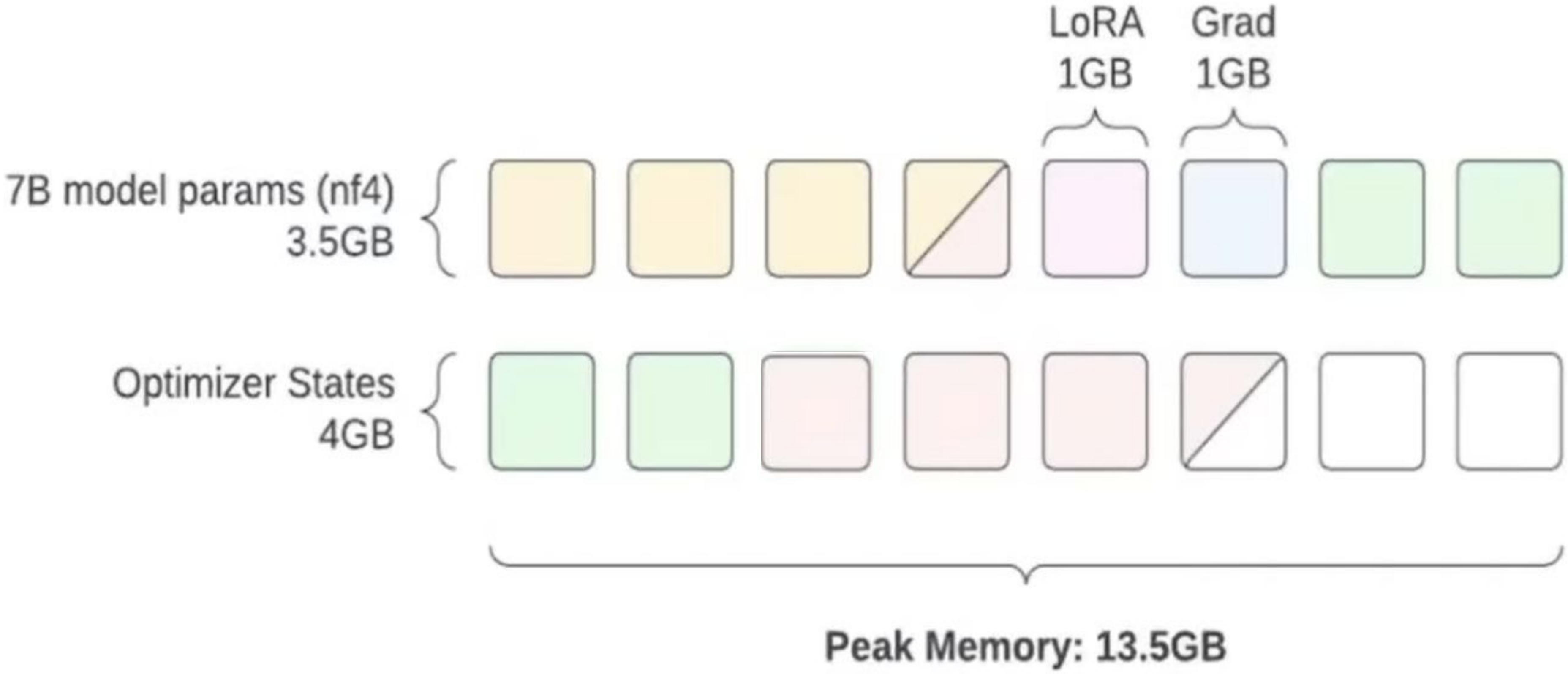
4-bit integers can represent 16 levels

-1.0, -0.8667, -0.7333, -0.6,

-0.4667, -0.3333, -0.2, -0.0667,

0.0667, 0.2, 0.3333, 0.4667,

0.6, 0.7333, 0.8667, 1.0


LoRA
1GB
Grad
1GB
7B model params (nf4)
3.5GB
Optimizer States
4GB
Peak Memory: 13.5GB

More to do

- Gradient Accumulation
- Paged Optimizers
- Double Quantization
- AdaLoRA
- LongLoRA

etc…

References

Finetuning LLMs with LoRA and QLoRA: Insights from Hundreds of Experiments - Lightning AI

https://blog.eleuther.ai/transformer-math/

https://developer.nvidia.com/blog/accelerating-ai-training-with-tf32-tensor-cores/

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *arXiv preprint arXiv:2305.14314* (2023).

Li, Yixiao, et al. "LoftQ: LoRA-Fine-Tuning-Aware Quantization for Large Language Models." *arXiv preprint arXiv:2310.08659* (2023).

Chen, Yukang, et al. "LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models." *arXiv preprint arXiv:2309.12307* (2023).