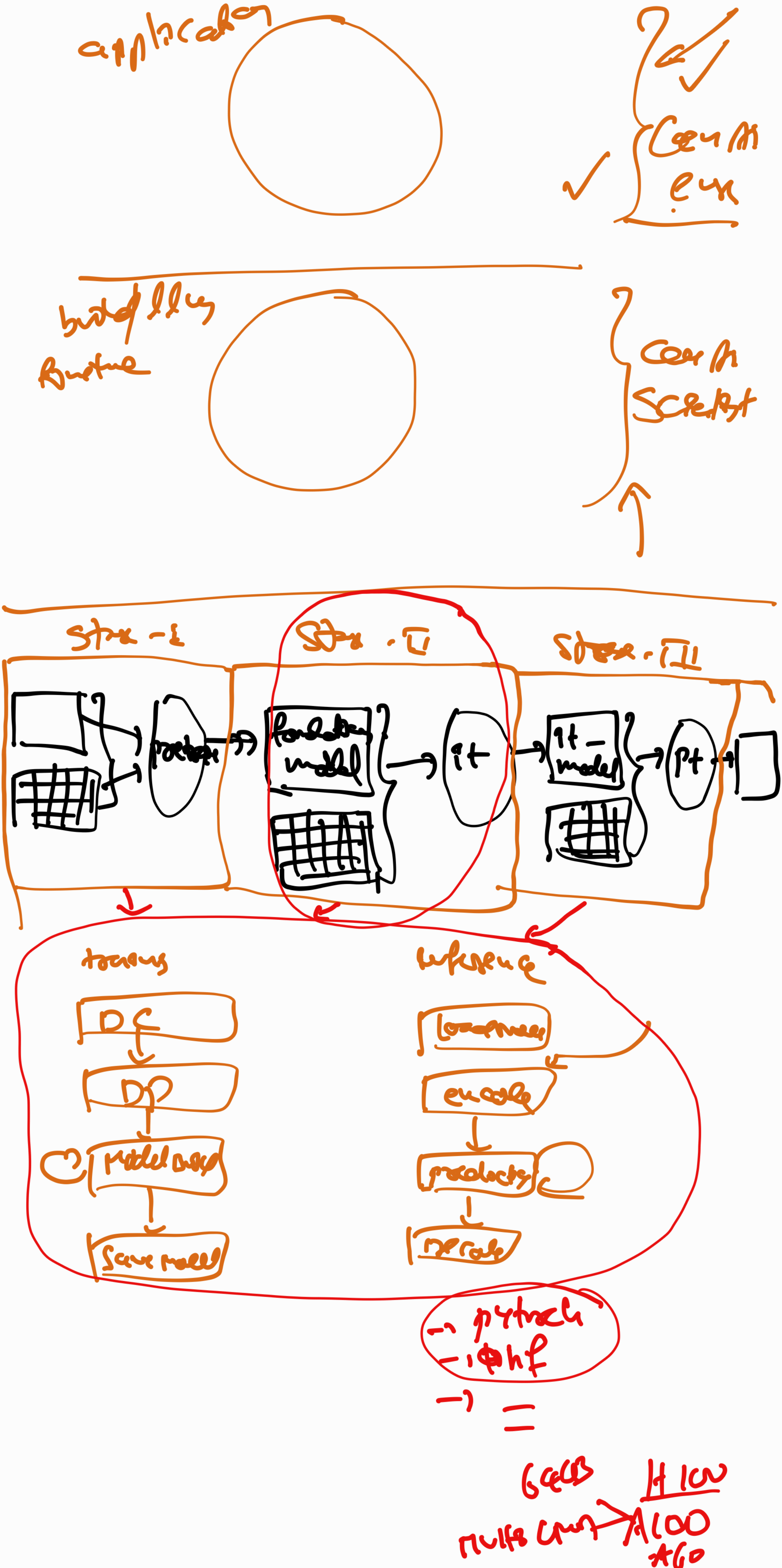
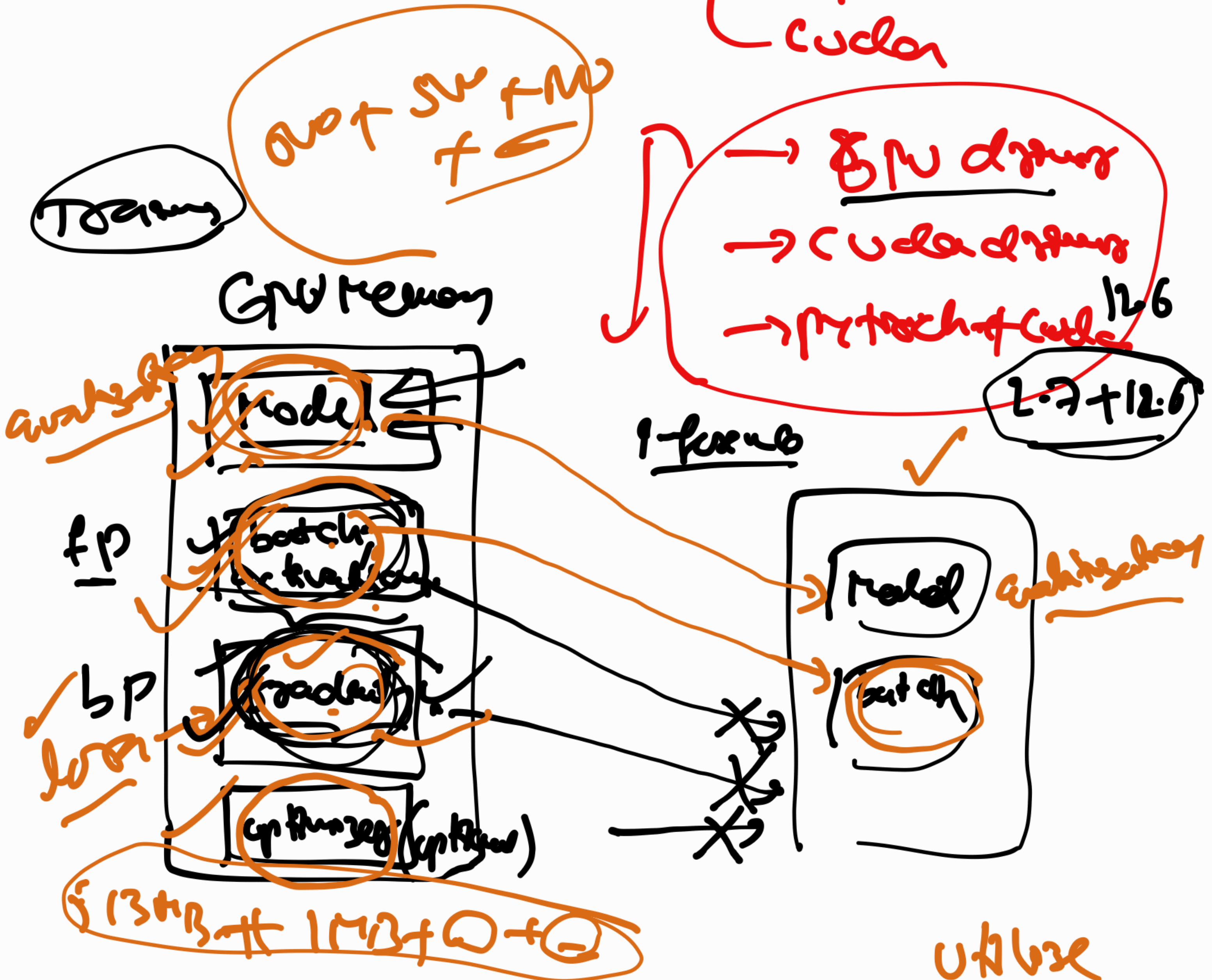

application
training
DP
inference
encode
pytorch
hf
H100
A100

Modern AI Series



Efficiency techniques to utilize / optimize GPU memory effectively

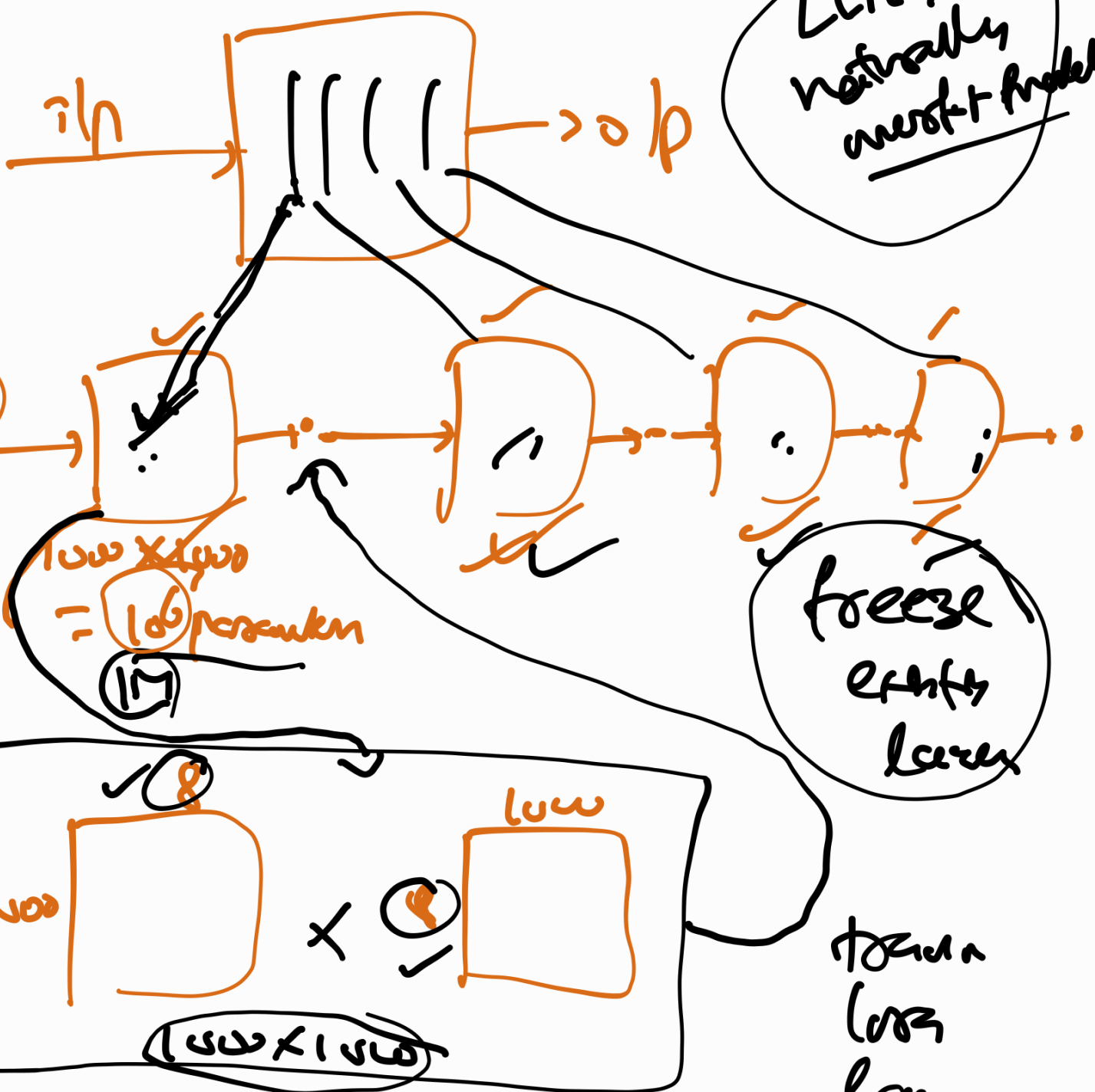
- LoRA
- Quantization
- Liger kernel
- Flash attention
- Caching

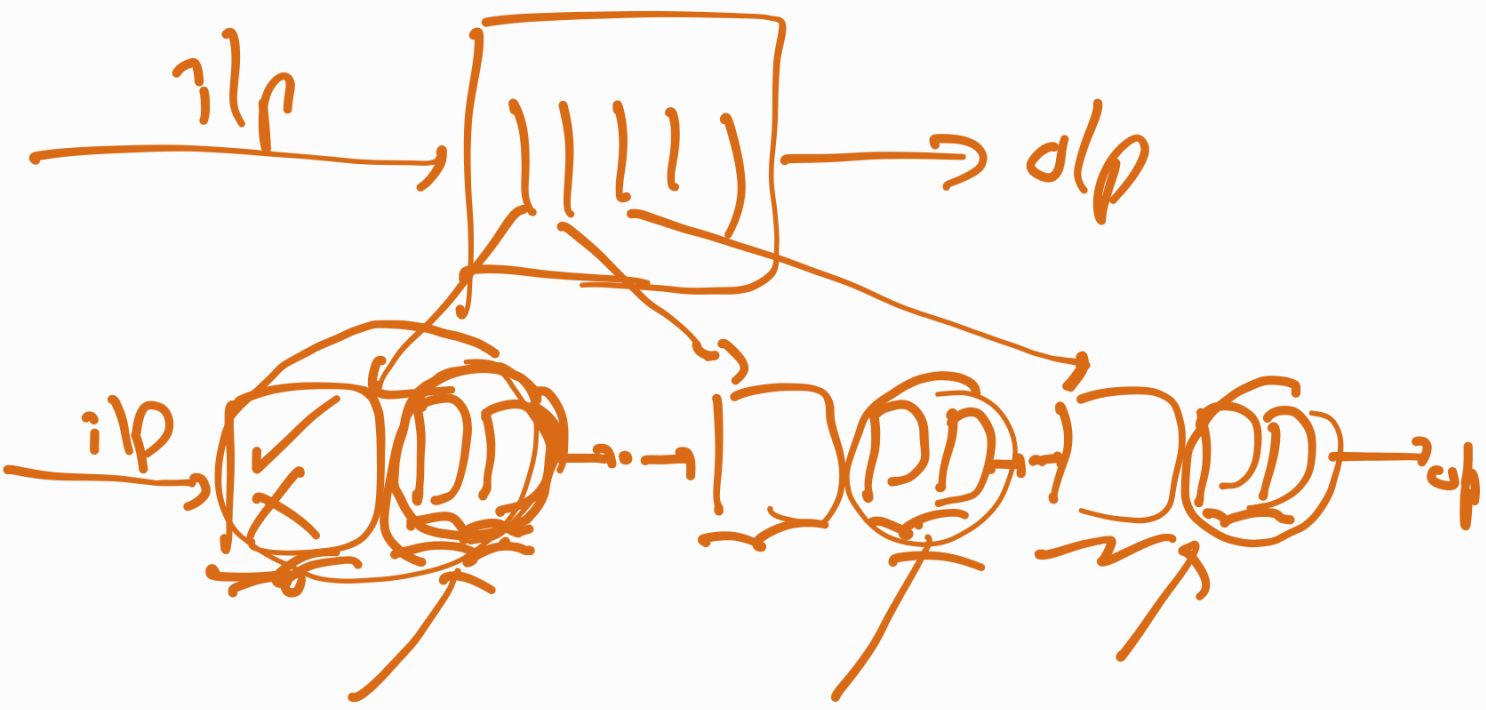
2023–25

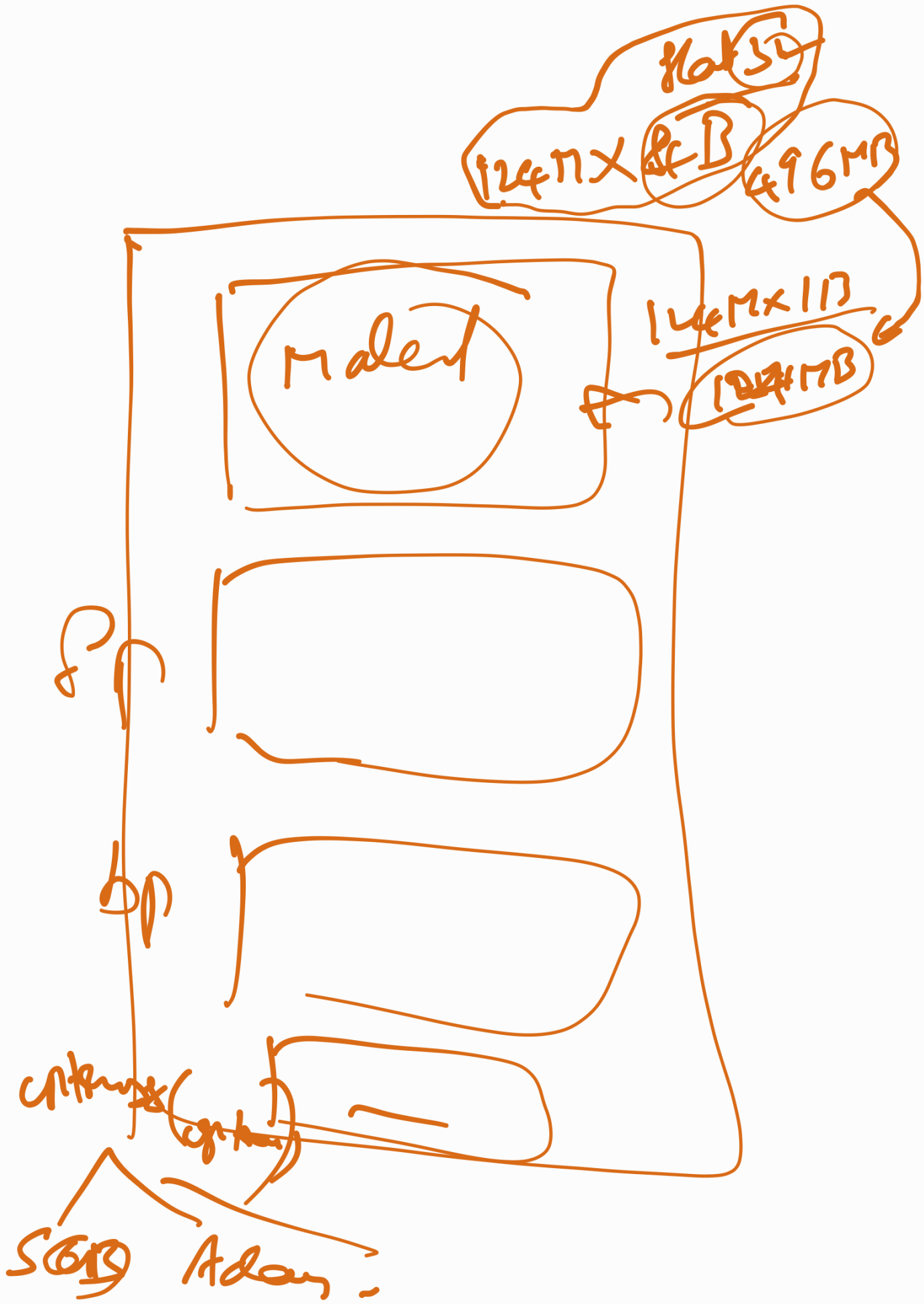
MoE

distillation

Low Rank Adaptor (LoRA)






model
SGD
Adam