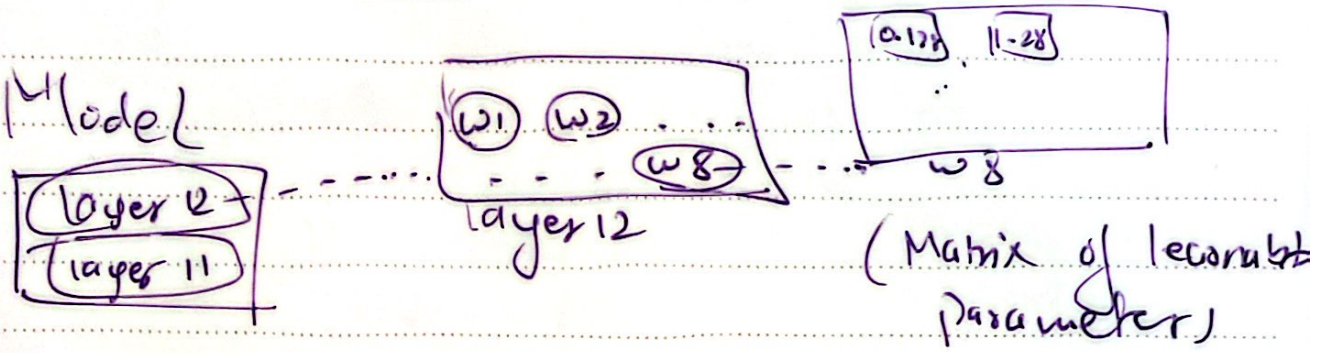
Lesson 3 Loading Models in different Datatypes



→ It is possible to inspect each parameter's datatype

→ casting using torch to load data as float16 instead of float 32, saves almost half the memory.

→ Look workdown for code L3

Lesson 4 ~~linear~~ Quantization Theory

Linear Quantization

① quantize → extreme values
→ highest value to int8 Max
lowest value to int8 min

191.6	-13.5	728.6
92.14	295.5	-184
0	684.6	245.5

→ 127

→ -128

Rest are mapped with (linear mapping)

some math formulae

final small matrix & s=scale, z= zero point

can get original with formulae
some errors

pip install transformers

pip install quanto

" " " torch

(look at Markdown) (L4)

pythia is in FP32, 400million param

→ 400×10^6 param

32 bits = 4 bytes

→ 8 bits = 1 byte

= $400 \times 10^6 \times 4$ bytes

= 1600×10^6 bytes

= 1600 megabytes

≈ 1.6 Gbytes

LLM quantization

- Many papers
- some require calibration step
- some ready to use
  - LLM.INT8 (only 8 bit)
  - linear quantization
  - QLORA (only 4-bit)
  - HQQ (upto 2-bit)

Llama 2 7B
- 28 GB storage in FP32 (32-bit precision)
- ~4 GB in 4 bit precision in GGUF format

→ The Bloke

Performance Degradation

Hugging face

Open LLM leaderboard