如何估算語言模型的訓練時間與計算量

許恆修

2025 十月

Overview

1. 被臨時叫進會議室的那一刻
2. 我們要怎麼估算訓練一個模型的「算力開銷」？
3. 從矩陣乘法到語言模型的計算公式
4. 案例研究：Llama3.1 405B怎麼算出70天的
5. 手算，還是電腦幫你算？

被臨時叫進會議室的那一刻

開場

- 匆匆忙忙，來滾帶爬
- 進到會議室中
- BOSS 輕鬆問句

「給你100張卡，訓練SLM，搞得起來嗎？」

被臨時叫進會議室的那一刻

場面話 vs. 腦內風暴

- 你微笑回答:「能！一定能！」

- 但腦袋狂飆:
  - 需要多少資料？
  - 需要多少人力？
  - 需要訓練多久？
  - 電費會不會爆？
  - 算力夠用嗎？

被臨時叫進會議室的那一刻

核心問題浮現

訓練一個模型，到底要花多少時間？

被臨時叫進會議室的那一刻

結論先行

$$\text{訓練時間} \approx \frac{6TP}{nX}$$

其中：

- T：語料中的總 Token 數量
- P：模型參數量（weights）
- n：GPU 卡數量
- X：每張 GPU 的 FLOPS 吞吐率

ref: Transformer Math 101

我們要怎麼估算訓練一個模型的算計開銷？

先學會 FLOPs 與 FLOPS

FLOPs：

- 指浮點運算次數, **Flo**ating **P**oint operation**s**
- 評估計算量的單位
- 能回答「這個模型要花多少力氣算完？」
- 在訓練時可估算 總運算量與訓練時間。
- 在部署時可用來比較 不同模型的運算開銷。
- 分析模型中哪個模組最吃算力（例如 Attention、FFN）

我們要怎麼估算訓練一個模型的算計開銷？

先學會 FLOPs 與 FLOPS

FLOPS：

- 指每秒浮點運算次數, Floating Point Operations per Second
- 描述硬體性能的指標

我們要怎麼估算訓練一個模型的算計開銷？

可以這樣記住：FLOPs 與 FLOPS

- FLOPs 是模型的「運算需求」
- FLOPS 是 GPU 的「算力」。

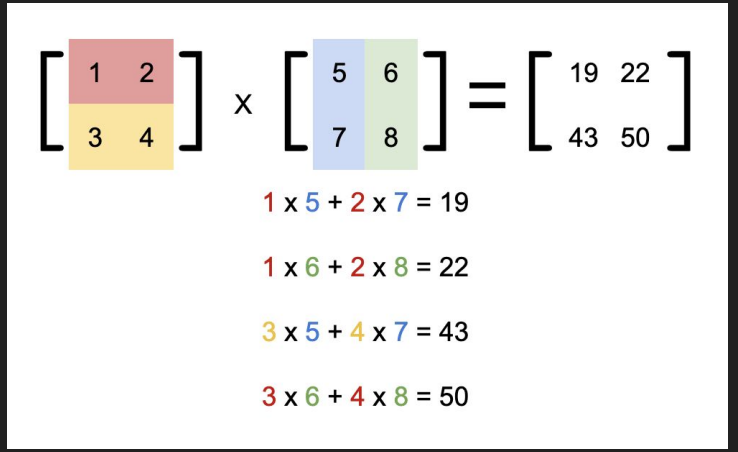
$$\frac{FLOPs}{FLOPS} = \frac{\text{運算需求}}{\text{硬體算力}}$$

把工作量除以體力，就是「訓練時間」。

從矩陣乘法到語言模型的計算公式

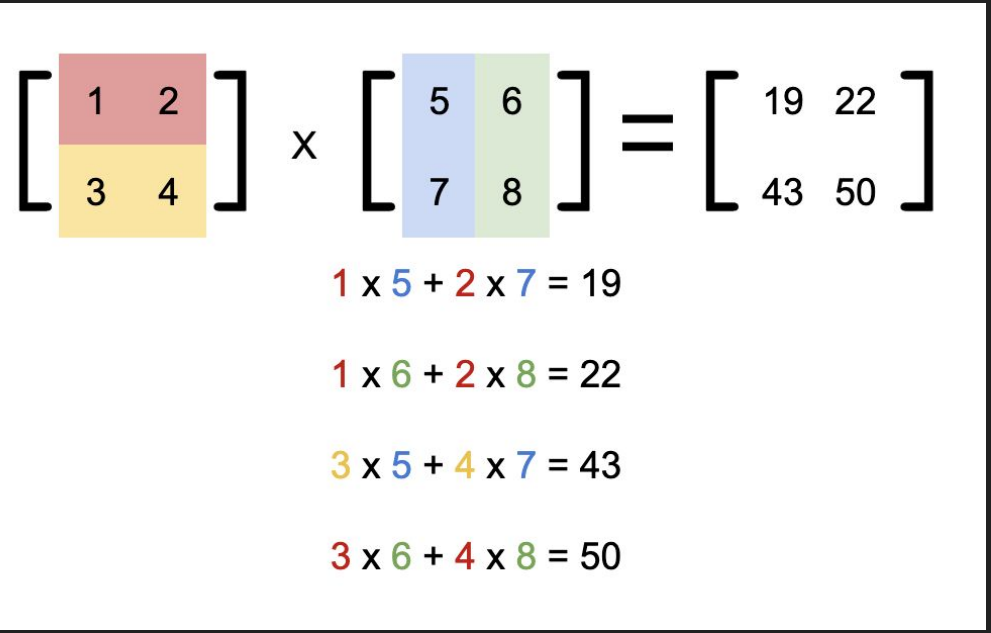
計算矩陣乘法的 FLOPs

- 深度學習背景下，許多運算都是矩陣乘法完成
- 尤其類神經網路，嚴重依賴矩陣成乘法
- 矩陣乘法涉及多個乘法和加法



從矩陣乘法到語言模型的計算公式

計算矩陣乘法的 FLOPs



- 考慮兩個矩陣 A 和 B
- 矩陣A 大小 a_1 * a_2
- 矩陣B 大小 b_1 * b_2
- 令 a_2 = b_1 = h
- C 的結果中反推 a_1 * b_2 元素中
- 每一個元素都需要經理 h 次乘法和 h-1 次加法
  - 化簡
  - **FLOPs = 2 X h X 輸出矩陣參數量**

從矩陣乘法到語言模型的計算公式

計算語言模型的 FLOPs

- 每一個 step，模型前向傳播擴展為
- *FLOPs = 2 X batch_size X sequence_length X 參數量*
- 又因反向傳播計算量為前向傳播的兩倍[backward-forward-FLOP-ratio]
- 一次step訓練的前向＋後向為 *(4+2) X batch_size X sequence_length X 參數量*
- 總共需 steps 做訓練所需 FLOPs
- 6 X *batch_size X sequence_length X 參數量*
- 其中 steps X batch_size X sequence_length = 語料 Token 總數量 T
- 最終得到 FLOPs = 6 X T X P

ref: scaling law, openai, 2020

案例研究：Llama3.1 405B 訓練時間估算

來看看 Meta AI 的 訓練時間多久

- 我們來看 Meta AI 的 Llama 3.1 405B ——
- 使用 15 兆 tokens、16,000 張 H100,
- 我們要試著「手算」它需要多少時間。

你覺得會是幾天？三個月？半年？

案例研究：Llama3.1 405B 訓練時間估算

$$\text{訓練時間} \approx \frac{6TP}{nX}$$

理論總模型計算量 FLOPs

6 X (405 x 10^9) X (15.6 x 10^12) = 3.8 X 10^25

	Training Data	Params	Input modalities	Output modalities	Context length	GQA	Token count	Knowled cutoff
Llama 3.1 (text only)	A new mix of publicly available online data.	8B	Multilingual Text	Multilingual Text and code	128k	Yes	15T+	Decembe 2023
		70B	Multilingual Text	Multilingual Text and code	128k	Yes		
		405B	Multilingual Text	Multilingual Text and code	128k	Yes		

案例研究：Llama3.1 405B 訓練時間估算

訓練時間 $\approx \frac{6TP}{nX}$

可用硬體算力 FLOPS

實驗中、H100 峰值為 400 TFLOPS，使用 16,000 張 H100

400 X 10^12 X (16 x 10^3) = 6.4 X 10^18

GPUs	TP	CP	PP	DP	Seq. Len.	Batch size/DP	Tokens/Batch	TFLOPs/GPU	BF16 MFU
8,192	8	1	16	64	8,192	32	16M	430	43%
16,384	8	1	16	128	8,192	16	16M	400	41%
16,384	8	16	16	8	131,072	16	16M	380	38%

Table 4 Scaling configurations and MFU for each stage of Llama 3 405B pre-training. See text and Figure 5 for descriptions of each type of parallelism.

案例研究：Llama3.1 405B 訓練時間估算

$$\text{訓練時間} \approx \frac{6TP}{nX}$$

算出訓練時間

將所需的 FLOPs 除以可用算力，再換成可理解的天數，最終可以得到

$$\text{訓練時間} = \frac{FLOPs}{FLOPS} = \frac{3.8 * 10^{25}}{6.4 * 10^{18} * 60 * 60 * 24} = 70\text{天}$$

案例研究：Llama3.1 405B 訓練時間估算

$$\text{訓練時間} \approx \frac{6TP}{nX}$$

對答案時間

看官方數據

- 總消耗 GPU 時數為 30.84M 小時。
- GPU小時計算方式為
- 單個 GPU 運算時間 X GPU 數量
- 30.84M 小時除 16,000 = 1927.5 小時
- 再除上 24 小時 = 80.3 天

	Training Time (GPU hours)	Training Power Consumption (W)	Training Location-Based Greenhouse Gas Emissions (tons CO2eq)
Llama 3.1 8B	1.46M	700	420
Llama 3.1 70B	7.0M	700	2,040
Llama 3.1 405B	30.84M	700	8,930

手算，還是電腦幫你算

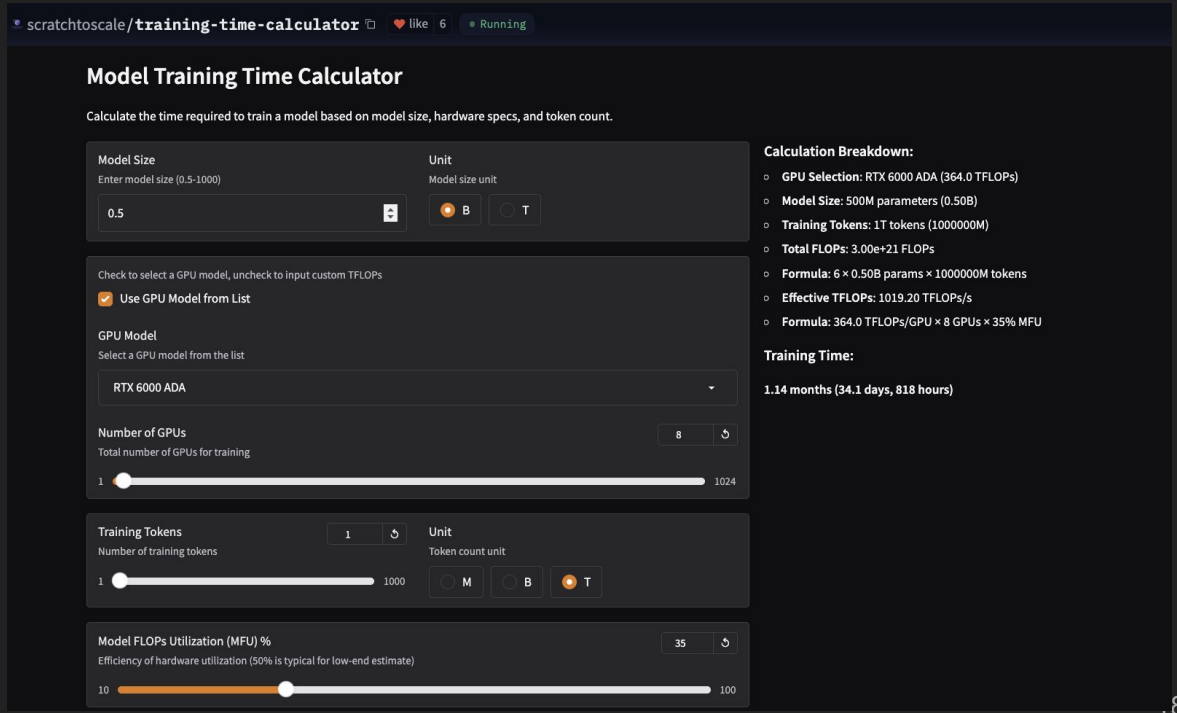
不用手算，直接用線上計算器估訓練時間。

開網址，輸入四件事

1. 參數量 P
2. 語料 Tokens T
3. GPU 數 n
4. 單卡 FLOPS X

立刻得到：

- 訓練時間



ref:scratchtoscale/training-time-calculator