

一條直線預測大模型

本課講 2020 年的一篇論文。OpenAI 的 Kaplan 和同事們發現：AI 變大以後會好多少，是可以提前算出來的。

適合 10 歲左右

約 40 分鐘

需要：鉛筆、尺子、15 張廢紙、1-3 位同伴

1 熱身：找規律

先填空。每一組都有自己的規律。

1、10、100、 3、30、300、 8、4、2、 64、32、16、

再看小明練寫一個難字。下面的數是我們編的：

寫了幾遍	1 遍	10 遍	100 遍	1000 遍
寫錯了幾筆	8 筆	4 筆	2 筆	<u> </u> 筆

你猜是 1 筆吧？規律是：**練習多 10 倍，錯誤就少一半**。發現了規律，沒練過的那一格也能猜出來。今天這篇論文，就是給 AI 找到了這樣的規律。

2 「越多越好」，到底好多少？

第 22 課我們發現：網絡越大，聽的錄音越多，本領就越強。可是還有一個問題沒回答：**多花 10 倍的錢，能好多少？**

2020 年，Kaplan 他們訓練了很多個「猜下一個字」的模型（第 1 課）。最大的有 15 億個旋鈕，比最小的多 100 多萬倍。旋鈕就是第 4 課講的電腦裡可以擰的數，大名叫**參數**。

怎麼比誰好？科學家給每次猜測打一個**錯誤分數**。猜對了，而且很有把握，分數就低。猜錯了，或者心裡沒底，分數就高。這個分數的大名叫**損失**，越低越好。他們改變下面 3 樣東西，看錯誤分數怎麼變：

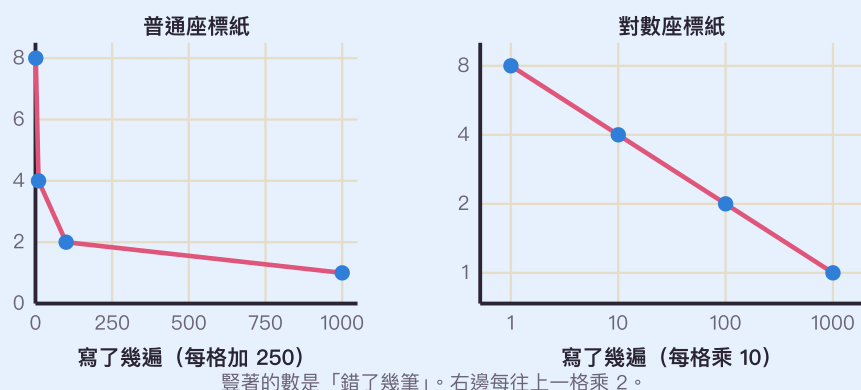
模型大小（腦袋有多大）
模型裡有多少個旋鈕。

數據（讀了多少書）
練習時讀了多少文字。

計算量（下了多少功夫）
練習時一共做了多少次算術。
旋鈕和書越多，算術越多。

3 特殊的座標紙：一格是 10 倍

把小明練字的數畫成圖，左右各畫一次：



普通座標紙上，1、10、100 擠在最左邊，線像一個急轉彎的滑梯，很難往後猜。右邊的座標紙不一樣。**每往右一格，數就乘 10**：1、10、100、1000。這種座標叫**對數座標**。同樣的 4 個點，在這裡排成了一**條直線**。有了直線，用尺子往後一延長，就能猜下一個點。

Kaplan 他們把錯誤分數畫在對數座標紙上，3 樣東西**都幾乎是直線**。一樣東西每多 10 倍，另一樣就按同樣的比例變化，這種規律叫**幂律**。AI 的這類規律合起來叫**規模定律**，「規模」就是大小。

真實的變化比小明慢得多。假設錯誤分數原來是 100 分，把一樣東西變成 10 倍：

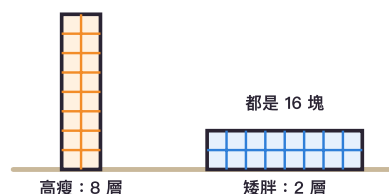
旋鈕 $\times 10$
約 84 分

書 $\times 10$
約 80 分

計算量 $\times 10$ (花得最好時)
約 89 分

每次只少十幾分，好像不多。可是旋鈕連著 4 次變成 10 倍，也就是多 1 萬倍，錯誤分數就降到大約一半。注意：只加一樣，另外兩樣也得夠用。就像做蛋糕，雞蛋不夠，麵粉再多也沒用。

4 高瘦還是矮胖，不太要緊



模型像一座樓，一層一層地疊起來。層數多就是「高」，每層的旋鈕多就是「胖」。用同樣多的旋鈕，可以搭一座高瘦的樓，也可以搭一座矮胖的樓。哪種更好？

論文發現：旋鈕總數一樣時，層數和每層寬度變來變去，錯誤分數**只差幾個百分點**。論文自己總結說：表現主要看規模有多大，跟形狀關係不大。當然，太極端的形狀不算，比如只有 1 層。所以科學家**先定旋鈕總數**就行。

5 先做小樣，再開工

做衣服的師傅先用便宜的布做小樣，合適了再剪貴布料。訓練大模型要用很多顯卡練很久（第 8 課、第 24 課），也很貴。規模定律讓科學家也能先做小樣：



花錢之前，心裡就有數了。論文還算了一筆賬：**錢有限時，怎麼分最划算？**

- 大部分錢花在把模型變大上。大模型學得更快，達到同樣的分數，讀的書更少。所以不用練到學不動，提前停下更划算。
- 書也要跟著多讀一些。書太少，大模型會死記硬背（第 4 課）。論文算出：模型放大 8 倍，書大約多讀 5 倍就夠了。

後來的人修正了這個結論

2022 年，DeepMind 公司的 Chinchilla 研究認為：**模型和書應該按同樣的比例一起加**，書不能少讀。Chinchilla 模型有 700 億個旋鈕，只有 Gopher 的四分之一，可是讀的書多得多，結果更厲害。原因之一：2020 年的實驗裡，所有模型都用同一份練習計劃，沒有按練多久來調整。科學就是這樣，後人檢查前人的結論，發現問題就改正。

6 這和今天有什麼關係

同一年，OpenAI 發佈了 GPT-3。它有 1750 億個旋鈕，讀的書相對少，正好照著這篇論文的建議來做。第二年，Kaplan 和第 22 課的 Dario Amodei 等人一起創辦了 Anthropic 公司，就是做聊天 AI Claude 的公司。

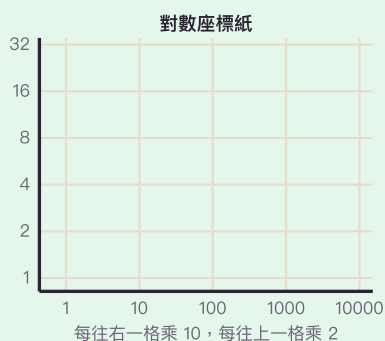
今天，各家公司訓練大模型以前，都會**先用小模型畫線**。這個辦法活下來了，具體怎麼分錢卻改了好幾次。也要誠實地說：規模定律是從實驗裡量出來的，不是一定成立的自然法則。直線畫得太遠，預測可能就不準了。現在網上好的文字快被讀完了，書不夠讀，也成了新的難題。

7 動手：對數座標紙

玩法（1 個人，約 10 分鐘）

下面兩組數是編的。旋鈕數和書的單位都是「萬」。

A 組：旋鈕數	1	10	100	1000	10000
錯誤分數	32	16	8	預測：_____	預測：_____
B 組：讀了幾萬本書		10	100	1000	
錯誤分數		32	8	預測：_____	



1. 把 A 組的 3 個點畫在左邊的對數座標紙上。2. 用尺子連線，往右延長，讀出兩個預測值。3. 換一種顏色畫 B 組，連線延長，預測 1000 處的分數。

你會發現：點排成了直線，沿著直線就能預測。B 組的線更陡，說明讀書多 10 倍，分數降得更多。真實論文裡也是這樣，只是差得小：書 80 分，旋鈕 84 分。

8 動手：小樣預測大貨

玩法（2—4 個人一組，約 20 分鐘）

1. 用 1 張廢紙搭紙塔，只能折、卷、撕。限時 3 分鐘，塔站穩 5 秒後量高度。再用 2 張、4 張各搭一座。2. 把結果畫在圖上，刻度自己標。連線延長到 8 張，讀出預測。3. 真用 8 張搭一次，看準不準。

用幾張紙	塔高（釐米）
1 張	
2 張	
4 張	
8 張：預測	
8 張：實際	



你會發現：預測多半對不上，大致高度卻能猜到。點太少，線容易畫歪。論文用了很多個模型，線更穩。

9 詞語小卡片

損失（錯誤分數）

給 AI 每次猜測打的分。猜得越離譜，分越高。越低越好。

對數座標

每往前一格，數就乘同一個倍數，比如 1、10、100、1000。

冪律

一樣東西每多 10 倍，另一樣就按同樣的比例變化。在對數座標紙上是直線。

規模定律

模型越大、書越多、算得越多，錯誤分數就按冪律變少。

10 想一想

1. 為什麼要用對數座標紙，不用普通座標紙？用你自己的話說一說。

2. 你有一筆錢訓練 AI。按 2020 年的論文，錢主要花在哪？按 2022 年的 Chinchilla 研究呢？

3. 紙塔用 8 張紙能預測，用 1000 張紙還能按同一條線預測嗎？為什麼？

一句話記住：模型越大、書越多、算得越多，AI 就按規律變好，先用小模型畫一條線，就能預測大模型。

給家長和老師：本課講的是 Jared Kaplan、Sam McCandlish、Tom Henighan 等人（OpenAI）2020 年的論文 [Scaling Laws for Neural Language Models](#)。非嵌入參數 768 到 15 億的 Transformer 語言模型，測試交叉熵損失分別隨參數量 N 、數據量 D 、算力 C 按冪律下降（另兩項不成瓶頸時，10 倍約降到 84%、80%、89%）；形狀影響小；過擬合取決於 $N^{0.74}/D$ （模型 8 倍、數據約 5 倍）；固定算力應優先放大模型並提前停止。2022 年 Chinchilla（DeepMind）修正為參數與數據同比例增長。

「錯誤分數」是交叉熵損失的通俗說法，「100 分」只表示比例；「相差 100 多萬倍」指非嵌入參數。小明練字、兩張座標圖、A/B 兩組數據和紙塔都是示意，示意數據比真實情況降得快得多。活動 7 答案：A 組 4 和 2，B 組 2。想深入可看本站的 [第 25 天中文導讀](#)，代碼見 [30u30 第 25 天目錄](#)（英文）。