

一条直线预测大模型

本课讲 2020 年的一篇论文。OpenAI 的 Kaplan 和同事们发现：AI 变大以后会好多少，是可以提前算出来的。

适合 10 岁左右

约 40 分钟

需要：铅笔、尺子、15 张废纸、1—3 位同伴

1 热身：找规律

先填空。每一组都有自己的规律。

1、10、100、

3、30、300、

8、4、2、_____

64、32、16、

再看小明练写一个难字。下面的数是我们编的：

写了几遍	1 遍	10 遍	100 遍	1000 遍
写错了几笔	8 笔	4 笔	2 笔	_____ 笔

你猜是 1 笔吧？规律是：**练习多 10 倍，错误就少一半**。发现了规律，没练过的那一格也能猜出来。今天这篇论文，就是给 AI 找到了这样的规律。

2 “越多越好”，到底好多少？

第 22 课我们发现：网络越大，听的录音越多，本领就越强。可是还有一个问题没回答：**多花 10 倍的钱，能好多少？**

2020 年，Kaplan 他们训练了很多个“猜下一个字”的模型（第 1 课）。最大的有 15 亿个旋钮，比最小的多 100 多万倍。旋钮就是第 4 课讲的电脑里可以拧的数，大名叫**参数**。

怎么比谁好？科学家给每次猜测打一个**错误分数**。猜对了，而且很有把握，分数就低。猜错了，或者心里没底，分数就高。这个分数的大名叫**损失**，越低越好。他们改变下面 3 样东西，看错误分数怎么变：

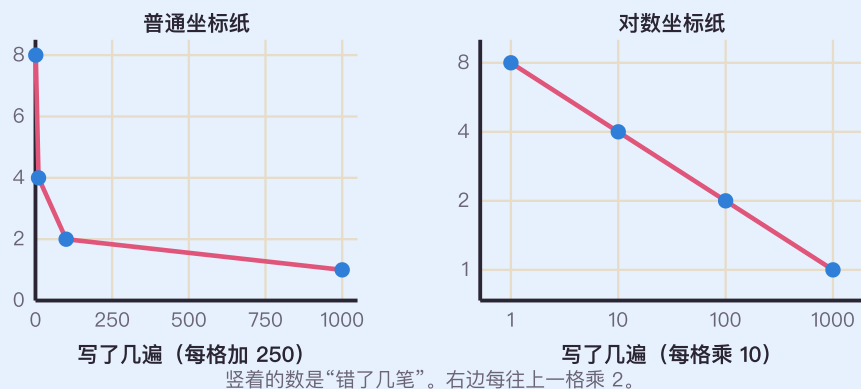
模型大小（脑袋有多大）
模型里有多少个旋钮。

数据（读了多少书）
练习时读了多少文字。

计算量（下了多少功夫）
练习时一共做了多少次算术。
旋钮和书越多，算术越多。

3 特殊的坐标纸：一格是 10 倍

把小明练字的数画成图，左右各画一次：



普通坐标纸上，1、10、100 挤在最左边，线像一个急转弯的滑梯，很难往后猜。右边的坐标纸不一样。**每往右一格，数就乘 10**：1、10、100、1000。这种坐标叫**对数坐标**。同样的 4 个点，在这里排成了一条直线。有了直线，用尺子往后一延长，就能猜下一个点。

Kaplan 他们把错误分数画在对数坐标纸上，3 样东西都几乎是直线。一样东西每多 10 倍，另一样就按同样的比例变化，这种规律叫**幂律**。AI 的这类规律合起来叫**规模定律**，“规模”就是大小。

真实的变化比小明慢得多。假设错误分数原来是 100 分，把一样东西变成 10 倍：

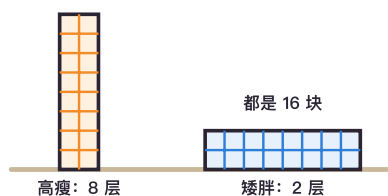
旋钮 $\times 10$
约 84 分

书 $\times 10$
约 80 分

计算量 $\times 10$ （花得最好时）
约 89 分

每次只少十几分，好像不多。可是旋钮连着 4 次变成 10 倍，也就是多 1 万倍，错误分数就降到大约一半。注意：只加一样，另外两样也得够用。就像做蛋糕，鸡蛋不够，面粉再多也没用。

4 高瘦还是矮胖，不太要紧



模型像一座楼，一层一层地叠起来。层数多就是“高”，每层的旋钮多就是“胖”。用同样多的旋钮，可以搭一座高瘦的楼，也可以搭一座矮胖的楼。哪种更好？

论文发现：旋钮总数一样时，层数和每层宽度变来变去，错误分数**只差几个百分点**。论文自己总结说：表现主要看规模有多大，跟形状关系不大。当然，太极端的形状不算，比如只有 1 层。所以科学家**先定旋钮总数**就行。

5 先做小样，再开工

做衣服的师傅先用便宜的布做小样，合适了再剪贵布料。训练大模型要用很多显卡练很久（第 8 课、第 24 课），也很贵。规模定律让科学家也能先做小样：



花钱之前，心里就有数了。论文还算了一笔账：**钱有限时，怎么分最划算？**

- 大部分钱花在把模型变大上。大模型学得更快，达到同样的分数，读的书更少。所以不用练到学不动，提前停下更划算。
- 书也要跟着多读一些。书太少，大模型会死记硬背（第 4 课）。论文算出：模型放大 8 倍，书大约多读 5 倍就够了。

后来的人修正了这个结论

2022 年，DeepMind 公司的 Chinchilla 研究认为：**模型和书应该按同样的比例一起加**，书不能少读。Chinchilla 模型有 700 亿个旋钮，只有 Gopher 的四分之一，可是读的书多得多，结果更厉害。原因之一：2020 年的实验里，所有模型都用同一份练习计划，没有按练多久来调整。科学就是这样，后人检查前人的结论，发现问题就改正。

6 这和今天有什么关系

同一年，OpenAI 发布了 GPT-3。它有 1750 亿个旋钮，读的书相对少，正好照着这篇论文的建议来做。第二年，Kaplan 和第 22 课的 Dario Amodei 等人一起创办了 Anthropic 公司，就是做聊天 AI Claude 的公司。

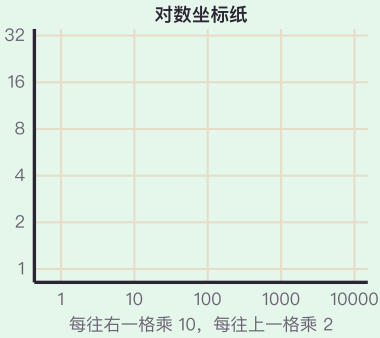
今天，各家公司训练大模型以前，都会**先用小模型画线**。这个办法活下来了，具体怎么分钱却改了好几次。也要诚实地说：规模定律是从实验里量出来的，不是一定成立的自然法则。直线画得太远，预测可能就不准了。现在网上好的文字快被读完了，书不够读，也成了新的难题。

7 动手：对数坐标纸

玩法（1 个人，约 10 分钟）

下面两组数是编的。旋钮数和书的单位都是“万”。

A 组：旋钮数	1	10	100	1000	10000
错误分数	32	16	8	预测：_____	预测：_____
B 组：读了几万本书		10	100	1000	
错误分数		32	8	预测：_____	



1. 把 A 组的 3 个点画在左边的对数坐标纸上。2. 用尺子连线，往右延长，读出两个预测值。3. 换一种颜色画 B 组，连线延长，预测 1000 处的分数。

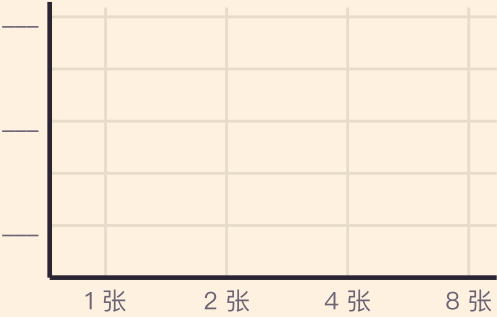
你会发现：点排成了直线，沿着直线就能预测。B 组的线更陡，说明读书多 10 倍，分数降得更多。真实论文里也是这样，只是差得小：书 80 分，旋钮 84 分。

8 动手：小样预测大货

玩法（2—4 个人一组，约 20 分钟）

1. 用 1 张废纸搭纸塔，只能折、卷、撕。限时 3 分钟，塔站稳 5 秒后量高度。再用 2 张、4 张各搭一座。2. 把结果画在图上，刻度自己标。连线延长到 8 张，读出预测。3. 真用 8 张搭一次，看准不准。

用几张纸	塔高（厘米）
1 张	
2 张	
4 张	
8 张：预测	
8 张：实际	



你会发现：预测多半对不上，大致高度却能猜到。点太少，线容易画歪。论文用了很多个模型，线更稳。

9 词语小卡片

损失（错误分数）

给 AI 每次猜测打的分。猜得越离谱，分越高。越低越好。

对数坐标

每往前一格，数就乘同一个倍数，比如 1、10、100、1000。

幂律

一样东西每多 10 倍，另一样就按同样的比例变化。在对数坐标纸上直线。

规模定律

模型越大、书越多、算得越多，错误分数就按幂律变少。

10 想一想

1. 为什么要用对数坐标纸，不用普通坐标纸？用你自己的话说一说。

2. 你有一笔钱训练 AI。按 2020 年的论文，钱主要花在哪？按 2022 年的 Chinchilla 研究呢？

3. 纸塔用 8 张纸能预测，用 1000 张纸还能按同一条线预测吗？为什么？

一句话记住：模型越大、书越多、算得越多，AI 就按规律变好，先用小模型画一条线，就能预测大模型。

给家长和老师：本课讲的是 Jared Kaplan、Sam McCandlish、Tom Henighan 等人（OpenAI）2020 年的论文 [Scaling Laws for Neural Language Models](#)。非嵌入参数 768 到 15 亿的 Transformer 语言模型，测试交叉熵损失分别随参数量 N 、数据量 D 、算力 C 按幂律下降（另两项不成瓶颈时，10 倍约降到 84%、80%、89%）；形状影响小；过拟合取决于 $N^{0.74}/D$ （模型 8 倍、数据约 5 倍）；固定算力应优先放大模型并提前停止。2022 年 Chinchilla（DeepMind）修正为参数与数据同比例增长。

“错误分数”是交叉熵损失的通俗说法，“100 分”只表示比例；“相差 100 多万倍”指非嵌入参数。小明练字、两张坐标图、A/B 两组数据和纸塔都是示意，示意数据比真实情况降得快得多。活动 7 答案：A 组 4 和 2，B 组 2。想深入可看本站的 [第 25 天中文导读](#)，代码见 [30u30 第 25 天目录](#)（英文）。