

每个字都抬头看一看

本课讲 2017 年的一篇论文《注意力就是你需要的全部》(Attention Is All You Need)。Vaswani 等 8 位科学家做出了一个新结构，叫 Transformer。今天的聊天 AI，几乎都是用它搭起来的。

适合 10 岁左右

约 40 分钟

需要：铅笔、空白卡片、一把豆子、几位同伴

1 热身：“它”是谁？

读每句话。在横线上写出“它”或“他”指的是谁。

① 小猫追着球跑，因为它很开心。它 = _____

② 小猫追着球跑，因为它滚得很快。它 = _____

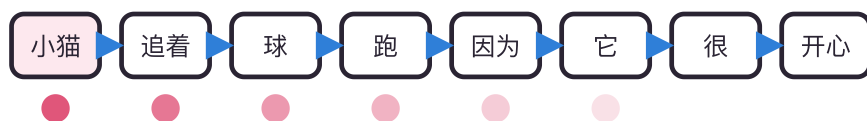
③ 小狗把骨头埋在树下，因为它想明天再吃。它 = _____

④ 姐姐把蛋糕分给弟弟，因为他饿了。他 = _____

①和②只差几个字，“它”就换了人。你是怎么知道的？你的眼睛**跳回前面**，看了看“小猫”和“球”。“开心”配小猫，“滚”配球。要读懂一个字，常常得回头看看别的字。

2 老办法：排队读，又慢又容易忘

第 1 课讲过**循环神经网络**。它带着一个“小本子”，一个字一个字往下读。第 2 课的 LSTM 给小本子加了“三扇门”，记得更久了。2017 年以前，电脑做翻译主要靠它们。它们有同一个规矩：**排队读**。

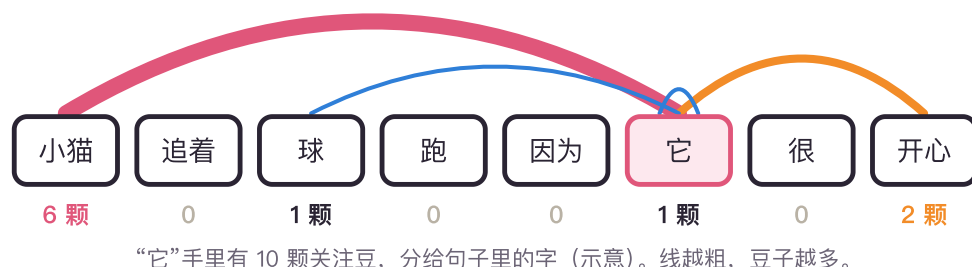


粉色圆点：小本子里“小猫”的线索，一步一步变淡（示意）

1. 慢。第 6 个字要等第 5 个字读完。就像食堂只开一个窗口，后面的人只能干等。
2. 隔得远就容易忘。读到“它”时，“小猫”已经是 5 步以前的事了。它的线索靠小本子一步一步往后带，路上会变淡。

3 新办法：每个字同时抬头看

Vaswani 他们想：为什么非要排队呢？他们的新结构叫 **Transformer**，英文意思是“变换器”。句子里**每个字同时抬头**，看一看所有的字，再决定自己该多关注谁。这件事叫**注意力**。看的是同一句话里的字，所以也叫**自注意力**。



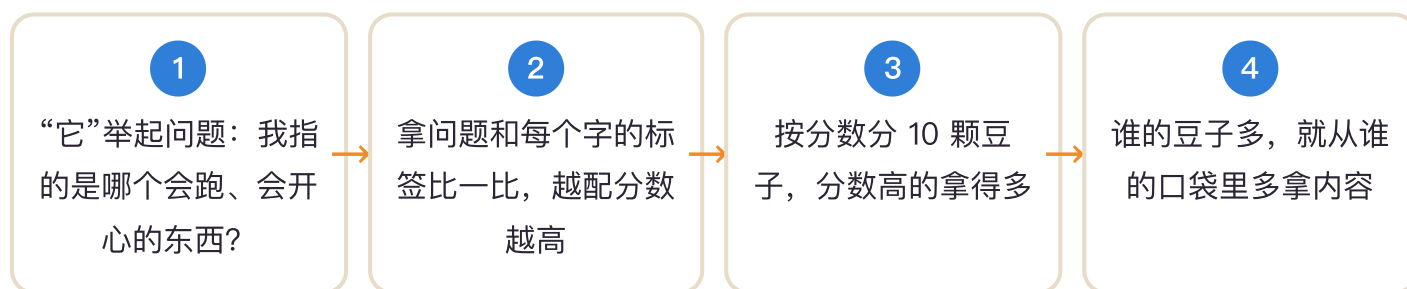
把“关注”想成**豆子**。每个字手里都有 10 颗关注豆，要全部分出去，也可以留给自己。“它”最想知道“我指的是谁”，所以把 6 颗给了“小猫”。

不只是“它”在分豆子：每个字都在同一时间分自己的豆子。

多远都只隔一步：“它”不用等小本子把“小猫”带过来，一抬头就看到了。

4 提问、标签、内容

豆子怎么分？Transformer 给每个字准备了三样东西：手里举着一个**提问**（我想找什么样的字），身上贴着一张**标签**（我是什么样的字），口袋里装着**内容**（被关注时分给别人的信息）。英文叫 Query、Key、Value。拿“它”走一遍（示意）：



“小猫”的标签是“我是会跑的小动物”，很配，拿 6 颗。“开心”的标签是“我说的是心情”，有一点配，拿 2 颗。“球”的标签是“我是会滚的东西”，不太配，只拿 1 颗。“它”把拿到的内容混在一起，变成自己新的意思。新意思里有很多“小猫”的成分，电脑就“明白”了：这里的“它”是小猫。

标签和内容为什么分开？ 像图书馆的书：书名是标签，用来找书；书里的故事才是内容。

问题和标签是谁写的？ 没有人写。它们其实是一串串数字，不是上面那些话。电脑在训练中自己学会了该问什么、该贴什么。

5 再加三个零件：小组、座位号、近路

零件一：好几组同时看（多头注意力）。一句话里的关系很多。Transformer 分成好几个**小组**同时看，每组从自己的角度分豆子，最后拼在一起。论文用了 **8 组**，一组叫一个“头”。比如“小猫在花园里追球”里的“追”（示意）：一组关心谁做了什么，豆子给“小猫”和“球”；另一组关心在哪里，豆子给“花园”。

零件二：座位号（位置编码）。大家同时看，就没人排队了。麻烦来了：电脑分不出谁前谁后。“狗咬人”和“人咬狗”在它看来一模一样。所以每个字进门时都要贴一个**座位号**，这叫**位置编码**。论文里的座位号是一串特别的数，每个座位都不一样。

零件三：每一层都有近路。Transformer 也是一层一层叠起来的。每一层先让大家**互相看**，再让每个字**自己想一想**。每一步旁边都有第 9 课讲过的**近路**，这一步只加上“改动”。论文做的是翻译：读原句的部分和写译文的部分，各叠了 6 层。

6 结果：同时算，显卡跑得飞快

不排队有个大好处：**所有字可以同时算**。第 8 课讲过，**显卡**像一整个班的同学，每人同时做一道小题。排队读的网络只能一步一步来，显卡大多在干等。Transformer 让全班一起开工。

论文拿它考了两门翻译：英文翻德文、英文翻法文。两门的分数都**超过了当时最好的模型**，还练得快多了。大一点的模型用 8 块显卡练了 3 天半。小一点的模型只练了 12 个小时。练得快，科学家就敢把模型做得更大，让它读更多的书。

7 这和今天有什么关系

你听说过 ChatGPT 吗？它名字里的 GPT 是三个英文词的缩写，**最后那个 T 就是 Transformer**。今天几乎所有的聊天 AI 都建在 Transformer 上。它们做的还是第 1 课的**猜下一个字**，只是换成了会“抬头看”的 Transformer 来猜。这些年它换过一些零件，比如座位号的写法，但“提问、标签、内容”和“好几组同时看”基本原样留了下来。

它也有麻烦：每个字都要看所有的字。10 个字要看 $10 \times 10 = 100$ 次，100 个字就要看 10000 次。句子一长，要算的就多得吓人。科学家还在想办法让它省力。

第 14 课会一行一行把 Transformer 写出来。第 15 课会讲：注意力最早是怎么来的。

8 动手：全班同时看

玩法（8 个人，约 15 分钟，需要 9 张空白卡片、10 颗豆子、秒表）

- 1. 做字卡：小猫、追着、球、跑、因为、它、很、开心。每人举一张，按顺序站成一排。
- 2. 第 1 轮 排队传：第 1 人小声对下一人说“小猫”。后面每人把听到的和自己的卡合起来，挑最多 4 个字往下传。举“它”的人猜“它”是谁。计时。
- 3. 第 2 轮 同时看：喊“开始”后，每人同时指向和自己最有关系的卡。举“它”的人把 10 颗豆子分到卡上。计时。
- 4. 第 3 轮 换字：把“开心”换成“滚得快”，“它”重新分豆子。

	几秒	“它”指谁	小猫的豆	球的豆	其他的豆
第 1 轮 排队传					
第 2 轮 同时看					
第 3 轮 换字					

你会发现：排队传又慢，“小猫”还可能半路被挤掉。同时看，一下子就指好了。只换一个字，豆子就从“小猫”搬到了“球”。

9 动手：座位号游戏

玩法（2 到 4 个人，约 10 分钟，需要 9 张空白卡片）

- 1. A 偷偷做字卡，一张写一块，背面写座位号 1、2、3……：① 小狗 | 追着 | 小猫 | 跑 ② 哥哥 | 送给 | 妹妹 | 一本 | 书
- 2. A 把一句的卡打乱，正面朝上摆给 B。B 不许翻面，试着拼回原句，写在表里。
- 3. 再让 B 翻到背面，按座位号排好。2 句都做一遍。

	没有座位号时，B 拼出的句子	和原句一样？	有座位号，一样？
①			
②			

你会发现：没有座位号，B 能拼出通顺的句子，却不一定是原句。“小猫追着小狗跑”也通顺，意思却反了。有了座位号，一次就对。Transformer 同时看所有的字，所以也必须贴座位号。

10 词语小卡片

注意力（自注意力）

每个字同时看句子里所有的字，决定该多关注谁。

提问、标签、内容

问题和标签越配，就从那个字拿走越多内容。英文叫 Query、Key、Value。

多头注意力

分成好几个小组同时看，每组从自己的角度分豆子。论文用了 8 组。

位置编码（座位号）

给每个字贴上座位号，同时看也能分清谁在前、谁在后。

11 想一想

1. “妹妹把苹果递给奶奶，因为她够不着。”“她”把最多的豆子分给谁？

2. 如果让你给 Transformer 再加一个小组，你想让它专门看什么？比如时间、颜色，还是别的？

3. 每个字都要看所有的字。如果让聊天 AI 读完一整本书，你觉得会遇到什么麻烦？

一句话记住：Transformer 不排队，每个字同时看所有字，再决定关注谁。

给家长和老师：本课讲的是 Vaswani、Shazeer、Parmar 等 8 位作者 2017 年的论文 [Attention Is All You Need](#) (NeurIPS 2017)。Transformer 去掉循环和卷积，编码器、解码器各 6 层，用缩放点积注意力 (Q/K/V)、8 头注意力、正弦余弦位置编码，每个子层带残差连接和层归一化。WMT 2014 英德翻译 28.4 BLEU，比此前最好结果高 2 以上；英法 41.8 BLEU，8 块 P100 训练 3.5 天；基础模型训练 12 小时。

示意或简化：豆子数（对应和为 1 的 softmax 权重），标签和问题的文字（实为学到的向量），两个小组的分工（论文附录显示有的头像在追踪指代或句子结构，但不是人为分工），“小猫变淡”图，各处例句。“字”指 token，“自己想一想”指前馈层；没讲解码器掩码和编码器-解码器注意力。见 [第 13 天中文导读](#) 和 [30u30 第 13 天目录](#)（英文）。