

每個字都抬頭看一看

本課講 2017 年的一篇論文《注意力就是你需要的全部》(Attention Is All You Need)。Vaswani 等 8 位科學家做出了一個新結構，叫 Transformer。今天的聊天 AI，幾乎都是用它搭起來的。

適合 10 歲左右

約 40 分鐘

需要：鉛筆、空白卡片、一把豆子、幾位同伴

1 熱身：「它」是誰？

讀每句話。在橫線上寫出「它」或「他」指的是誰。

① 小貓追著球跑，因為它很開心。它 = _____

② 小貓追著球跑，因為它滾得很快。它 = _____

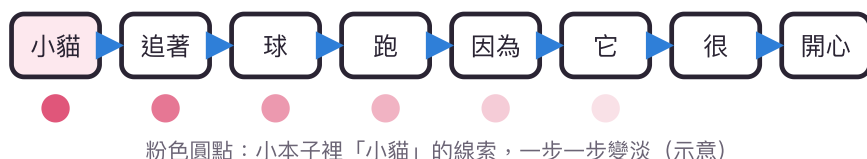
③ 小狗把骨頭埋在樹下，因為它想明天再吃。它 = _____

④ 姐姐把蛋糕分給弟弟，因為他餓了。他 = _____

①和②只差幾個字，「它」就換了人。你是怎麼知道的？你的眼睛**跳回前面**，看了看「小貓」和「球」。「開心」配小貓，「滾」配球。要讀懂一個字，常常得回頭看看別的字。

2 老辦法：排隊讀，又慢又容易忘

第 1 課講過**循環神經網絡**。它帶著一個「小本子」，一個字一個字往下讀。第 2 課的 LSTM 給小本子加了「三扇門」，記得更久了。2017 年以前，電腦做翻譯主要靠它們。它們有同一個規矩：**排隊讀**。

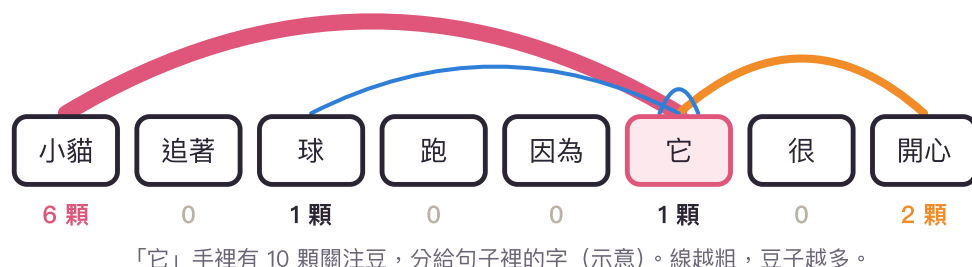


粉色圓點：小本子裡「小貓」的線索，一步一步變淡（示意）

1. 慢。第 6 個字要等第 5 個字讀完。就像食堂只開一個窗口，後面的人只能乾等。
2. 隔得遠就容易忘。讀到「它」時，「小貓」已經是 5 步以前的事了。它的線索靠小本子一步一步往後帶，路上會變淡。

3 新辦法：每個字同時抬頭看

Vaswani 他們想：為什麼非要排隊呢？他們的新結構叫 **Transformer**，英文意思是「變換器」。句子裡**每個字同時抬頭**，看一看所有的字，再決定自己該多關注誰。這件事叫**注意力**。看的是同一句話裡的字，所以也叫**自注意力**。



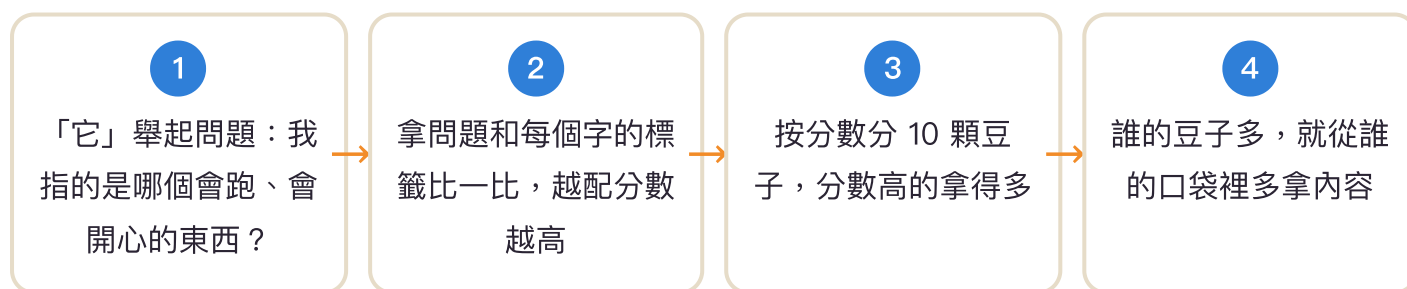
把「關注」想成**豆子**。每個字手裡都有 10 顆關注豆，要全部分出去，也可以留給自己。「它」最想知道「我指的是誰」，所以把 6 顆給了「小貓」。

不只是「它」在分豆子：每個字都在同一時間分自己的豆子。

多遠都只隔一步：「它」不用等小本子把「小貓」帶過來，一抬頭就看到了。

4 提問、標籤、內容

豆子怎麼分？Transformer 給每個字準備了三樣東西：手裡舉著一個**提問**（我想找什麼樣的字），身上貼著一張**標籤**（我是什麼樣的字），口袋裡裝著**內容**（被關注時分給別人的信息）。英文叫 Query、Key、Value。拿「它」走一遍（示意）：



「小貓」的標籤是「我是會跑的小動物」，很配，拿 6 顆。「開心」的標籤是「我說的是心情」，有一點配，拿 2 顆。「球」的標籤是「我是會滾的東西」，不太配，只拿 1 顆。「它」把拿到的內容混在一起，變成自己新的意思。新意思裡有很多「小貓」的成分，電腦就「明白」了：這裡的「它」是小貓。

標籤和內容為什麼分開？像圖書館的書：書名是標籤，用來找書；書裡的故事才是內容。

問題和標籤是誰寫的？沒有人寫。它們其實是一串串數字，不是上面那些話。電腦在訓練中自己學會了該問什麼、該貼什麼。

5 再加三個零件：小組、座位號、近路

零件一：好幾組同時看（多頭注意力）。一句話裡的關係很多。Transformer 分成好幾個**小組**同時看，每組從自己的角度分豆子，最後拼在一起。論文用了 **8 組**，一組叫一個「頭」。比如「小貓在花園裡追球」裡的「追」（示意）：一組關心誰做了什麼，豆子給「小貓」和「球」；另一組關心在哪裡，豆子給「花園」。

零件二：座位號（位置編碼）。大家同時看，就沒人排隊了。麻煩來了：電腦分不出誰前誰後。

「狗咬人」和「人咬狗」在它看來一模一樣。所以每個字進門時都要貼一個**座位號**，這叫**位置編碼**。論文裡的座位號是一串特別的數，每個座位都不一樣。

零件三：每一層都有近路。Transformer 也是一層一層疊起來的。每一層先讓大家**互相看**，再讓每個字**自己想一想**。每一步旁邊都有第 9 課講過的**近路**，這一步只加上「改動」。論文做的是翻譯：讀原句的部分和寫譯文的部分，各疊了 6 層。

6 結果：同時算，顯卡跑得飛快

不排隊有個大好處：**所有字可以同時算**。第 8 課講過，**顯卡**像一整個班的同學，每人同時做一道小題。排隊讀的網絡只能一步一步來，顯卡大多在乾等。Transformer 讓全班一起開工。

論文拿它考了兩門翻譯：英文翻德文、英文翻法文。兩門的分數都**超過了當時最好的模型**，還練得快多了。大一點的模型用 8 塊顯卡練了 3 天半。小一點的模型只練了 12 個小時。練得快，科學家就敢把模型做得更大，讓它讀更多的書。

7 這和今天有什麼關係

你聽說過 ChatGPT 嗎？它名字裡的 GPT 是三個英文詞的縮寫，**最後那個 T 就是 Transformer**。今天幾乎所有的聊天 AI 都建在 Transformer 上。它們做的還是第 1 課的**猜下一個字**，只是換成了會「抬頭看」的 Transformer 來猜。這些年它換過一些零件，比如座位號的寫法，但「提問、標籤、內容」和「好幾組同時看」基本原樣留了下來。

它也有麻煩：每個字都要看所有的字。10 個字要看 $10 \times 10 = 100$ 次，100 個字就要看 10000 次。句子一長，要算的就多得嚇人。科學家還在想辦法讓它省力。

第 14 課會一行一行把 Transformer 寫出來。第 15 課會講：注意力最早是怎麼來的。

8 動手：全班同時看

玩法（8 個人，約 15 分鐘，需要 9 張空白卡片、10 顆豆子、秒錶）

1. 做字卡：小貓、追著、球、跑、因為、它、很、開心。每人舉一張，按順序站成一排。
2. 第 1 輪 排隊傳：第 1 人小聲對下一人說「小貓」。後面每人把聽到的和自己的卡合起來，挑最多 4 個字往下傳。舉「它」的人猜「它」是誰。計時。
3. 第 2 輪 同時看：喊「開始」後，每人**同時**指向和自己最有關係的卡。舉「它」的人把 10 顆豆子分到卡上。計時。
4. 第 3 輪 換字：把「開心」換成「滾得快」，「它」重新分豆子。

	幾秒	「它」指誰	小貓的豆	球的豆	其他的豆
第 1 輪 排隊傳					
第 2 輪 同時看					
第 3 輪 換字					

你會發現：排隊傳又慢，「小貓」還可能半路被擠掉。同時看，一下子就指好了。只換一個字，豆子就從「小貓」搬到了「球」。

9 動手：座位號遊戲

玩法（2 到 4 個人，約 10 分鐘，需要 9 張空白卡片）

1. A 偷偷做字卡，一張寫一塊，**背面**寫座位號 1、2、3……：① 小狗 | 追著 | 小貓 | 跑 ② 哥哥 | 送給 | 妹妹 | 一本 | 書
2. A 把一句的卡打亂，正面朝上擺給 B。B 不許翻面，試著拼回原句，寫在表裡。
3. 再讓 B 翻到背面，按座位號排好。2 句都做一遍。

	沒有座位號時，B 拼出的句子	和原句一樣？	有座位號，一樣？
①			
②			

你會發現：沒有座位號，B 能拼出通順的句子，卻不一定是原句。「小貓追著小狗跑」也通順，意思卻反了。有了座位號，一次就對。Transformer 同時看所有的字，所以也必須貼座位號。

10 詞語小卡片

注意力（自注意力）

每個字同時看句子裡所有的字，決定該多關注誰。

提問、標籤、內容

問題和標籤越配，就從那個字拿走越多內容。英文叫 Query、Key、Value。

多頭注意力

分成好幾個小組同時看，每組從自己的角度分豆子。論文用了 8 組。

位置編碼（座位號）

給每個字貼上座位號，同時看也能分清誰在前、誰在後。

11 想一想

1. 「妹妹把蘋果遞給奶奶，因為她夠不著。」「她」把最多的豆子分給誰？

2. 如果讓你給 Transformer 再加一個小組，你想讓它專門看什麼？比如時間、顏色，還是別的？

3. 每個字都要看所有的字。如果讓聊天 AI 讀完一整本書，你覺得會遇到什麼麻煩？

一句話記住：Transformer 不排隊，每個字同時看所有字，再決定關注誰。

給家長和老師：本課講的是 Vaswani、Shazeer、Parmar 等 8 位作者 2017 年的論文 [Attention Is All You Need](#) (NeurIPS 2017)。Transformer 去掉循環和卷積，編碼器、解碼器各 6 層，用縮放點積注意力 (Q/K/V)、8 頭注意力、正弦餘弦位置編碼，每個子層帶殘差連接和層歸一化。WMT 2014 英德翻譯 28.4 BLEU，比此前最好結果高 2 以上；英法 41.8 BLEU，8 塊 P100 訓練 3.5 天；基礎模型訓練 12 小時。

示意或簡化：豆子數（對應和為 1 的 softmax 權重），標籤和問題的文字（實為學到的向量），兩個小組的分工（論文附錄顯示有的頭像在追蹤指代或句子結構，但不是人為分工），「小貓變淡」圖，各處例句。「字」指 token，「自己想一想」指前饋層；沒講解碼器掩碼和編碼器-解碼器注意力。見 [第 13 天中文導讀](#) 和 [30u30 第 13 天目錄](#)（英文）。