

請假請出萬千小隊

本課講 2014 年的一篇文章。Srivastava、Hinton、Krizhevsky、Sutskever 和 Salakhutdinov 五位科學家把「請假法」(Dropout) 講了個透：零件隨機請假，一個電腦大腦就能頂好多個。

適合 10 歲左右

約 40 分鐘

需要：鉛筆、一罐豆子、1 個骰子、幾位同伴

1 熱身：五個人猜書

5 位同學各看了一眼書架，猜有幾本書。老師數了一遍：一共 **52 本**。請算出每個人差了多少。

誰猜的	小東	小南	小西	小北	小中
猜了幾本	30	70	45	85	20
和 52 差多少					

把 5 個數加起來再除以 5，得到**平均數**：_____ 它和 52 差多少？_____

誰最準？請圈出來： 小東 小南 小西 小北 小中 平均數

算出來了嗎？最準的不是哪一個人，而是**大家的平均數**。這節課的秘密，就藏在這件事裡。

2 很多人一起猜，比一個人準

為什麼平均數這麼準？因為每個人錯的方向不一樣。有人猜多了，有人猜少了。一平均，多出來的和少掉的就**互相抵消**了一大半。

電腦也是這樣。一個練好的電腦大腦，叫一個**模型**。練出 5 個模型，它們死記硬背的地方各不一樣。遇到新題，把 5 個模型的答案平均一下，往往比任何一個單獨答都準。這個辦法叫**集成**，意思是「把好幾個模型的意見合在一起」。

可是，集成太貴了

一個大模型要用很多臺電腦練好幾天。練 100 個，就要花 100 倍的時間。回答一個問題，100 個模型也都要算一遍。有沒有辦法只練 1 個模型，卻得到「很多個模型一起投票」的好處？

3 每次請假，都是一支新小隊

第 3 課講過「請假法」：電腦練習時，每一次都隨機選一部分零件請假，剩下的零件照樣幹活。這篇文章換了一個角度看它：**每一次請假，到場的零件就組成了一支新的小隊。**

假設電腦大腦只有 3 個零件：甲、乙、丙。每個零件都有「來」和「不來」兩種情況，所以一共有 $2 \times 2 \times 2 = 8$ 種小隊：



藍色實心：到場。灰色虛線：請假。最後一種是全體請假，只是湊數用的。

每多 1 個零件，小隊的種數就**翻一倍**。10 個零件有 1024 種小隊，20 個零件有 100 多萬種。1000 個零件的小隊種數，寫出來是一個 302 位的數。

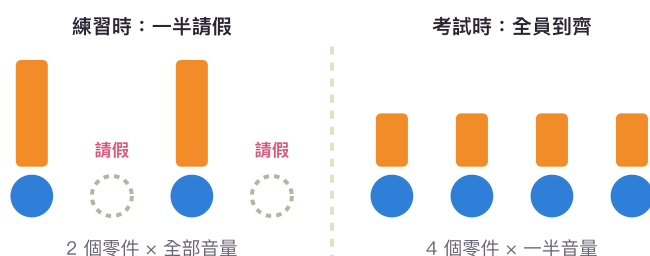
所有小隊共用同一套零件

全班每天隨機分成不同的小隊做任務。小隊天天換，同學還是那些同學。小華今天在一支小隊學會了算賬，明天到了別的小隊，本領也跟著帶過去。電腦裡，大部分小隊一次也沒輪上練習。可是零件是共用的，它們也跟著變好了。**練 1 個大腦，就等於練了千千萬萬支小隊。**

4 考試時：全員到齊，小聲說話

集成要讓所有小隊投票，再取平均。可是小隊多到數不清，一支一支去問，永遠問不完。科學家想了一個省事的辦法。

考試時誰也不請假，所有零件一起回答。可是每個零件要把音量調小，**按出勤比例打折**。為什麼？練習時，每個零件只有一半的時候到場。考試時人全來了，說話聲就比平時大了一倍。所以每個零件**只用一半的音量**說話，總音量就和練習時差不多了。



示意圖：橙色柱子是每個零件說話的音量。兩邊加起來的總音量一樣。

到場比例不同，折扣也不同。這篇文章裡，中間的零件常常一半請假，音量打 5 折。直接看圖片的那一排零件大約 8 成到場，音量就打 8 折。

這個省事的辦法靠譜嗎？

科學家隨機抽了大約 50 支小隊，分別回答再取平均。結果和「全員到齊、小聲說話」差不多。所以考試時只算 **1 次**，就近似得到了所有小隊的平均意見。

5 靈感：爸爸媽媽的基因怎麼搭配

科學家說，靈感之一來自生物。你身體裡有一套「說明書」，叫**基因**，它決定你的眼睛顏色、頭髮卷不卷。你的基因一半來自爸爸，一半來自媽媽。每個孩子拿到的組合都不一樣，所以兄弟姐妹長得像，又不完全一樣。

一個基因到了下一代，身邊的「搭檔基因」就換了一批。它沒法指望固定搭檔，**和各種搭檔都能合作，才算好基因**。

電腦裡的零件也會找固定搭檔。不請假時，兩個零件可能「抱團」：甲只在乙說某句話時才有用，兩個合起來死記練習捲上的巧合。這種抱團，科學家叫它**共適應**。請假法讓搭檔隨時可能不在，抱團就行不通了。就像基因要和各種搭檔合作，**每個零件也必須自己就有用**。

6 看、聽、讀，樣樣都管用

這篇文章試得特別多，科學家在 4 類任務上都用了請假法：

看圖	認手寫數字、認街景照片裡的門牌號、認照片裡是飛機還是貓
聽語音	聽一段說話聲，分出裡面一個一個的音
讀文字	讀一篇新聞，判斷它屬於哪一類
生物	預測基因的說明書會怎樣被剪開、拼起來

加上請假法以後，幾乎每個任務的錯誤都變少了。其中讀文字那個任務進步最小。請假法也有代價。每次只有一部分零件在練，學得慢一些。練習時間大約是原來的 2 到 3 倍。

7 這和今天有什麼關係

2012 年，看圖程序 AlexNet 用了請假法，在一場看圖大賽裡遙遙領先。這篇 2014 年的文章把道理講完整了，請假法從此成了常用工具。後來聊天 AI 用的大腦結構，最初的設計裡也用了它。第 3 課說過，最大的聊天 AI 開始大量讀書時常常不請假。可是「練習時加一點隨機」這個想法，今天很多訓練方法都在用。

「很多人一起猜」的道理你也用得上。問聊天 AI 難題時，可以多問幾次、多查幾處。答案不一致時，就要小心了。

8 動手：估豆子，看看誰最準

玩法（5 人以上，全班更好；約 15 分鐘；一個透明罐子裝豆子）

1. 老師先數好罐子裡的豆子，建議 100 到 300 顆，寫在紙上藏起來。
2. 每人看罐子 20 秒，**不許討論**，把猜的數偷偷寫下來。
3. 老師公佈真實數量。每人算出自己差了多少，填進下表。
4. 把所有人猜的數加起來，再除以人數，得到全班平均數（可以用計算器）。

名字	猜了幾顆	和真實數差多少	名字	猜了幾顆	和真實數差多少

真實數量：_____ 全班平均：_____ 平均數差了：_____ 比平均數更準的有 _____ 人
(人多就寫在背面)

你會發現：全班平均數常常比大多數人都準，因為猜多的和猜少的抵消了。大家都抄一個人，就沒法抵消。請假法的小隊也是到場成員不同，犯的錯才不同。

9 動手：隨機小隊記詞比賽

玩法（2 隊各 3 人加 1 位主持人；約 15 分鐘；1 個骰子）

1. 每隊 3 人編號 1—3。主持人每輪把表裡一行詞讀 2 遍，各隊 1 分鐘內寫下來。
2. **固定隊**：3 人都在。1 號只記動物，2 號只記水果，3 號只記交通工具。
3. **隨機隊**：練習前擲骰子，1 或 4 是 1 號請假，2 或 5 是 2 號，3 或 6 是 3 號。
4. 考試 1 兩隊全員到齊。考試 2 擲一次骰子，兩隊同號的人請假。請假的人捂住耳朵。

第幾輪	主持人讀的 9 個詞	固定隊	隨機隊
練習 1	貓、蘋果、汽車、狗、香蕉、火車、兔、葡萄、飛機		
練習 2	馬、西瓜、輪船、羊、梨、自行車、熊、桃、公交車		
考試 1	鴨、芒果、直升機、虎、櫻桃、出租車、蛇、菠蘿、馬車		
考試 2	鵝、荔枝、火箭、狼、椰子、帆船、鹿、李子、救護車		

右邊兩列填記住幾個詞。**你會發現：**考試 2 有人請假時，固定隊會整類地丟詞。隨機隊練慣了少一個人，每人都會記好幾類，丟得少。

10 詞語小卡片

集成

把好幾個模型的答案合在一起，比如取平均。通常比一個模型單獨答更準，但是很貴。

小隊（子網絡）

某一次練習時到場的零件。一個大腦裡有數不清的小隊，共用同一套零件。

全員到齊，小聲說話

考試時不請假，每個零件按出勤比例調小音量，近似所有小隊的平均意見。

共適應（抱團）

幾個零件互相依賴，一起死記練習捲上的巧合。隨機請假能拆散這種抱團。

11 想一想

1. 估豆子時，如果大家先討論、再一起寫一個數，平均數還會那麼準嗎？為什麼？

2. 考試時零件全來了，為什麼還要「小聲說話」？如果大家都用全部音量，會怎樣？

3. 你的班級或球隊裡，有沒有「只靠某一個人」的事？他請假了，大家怎樣頂上？

一句話記住：隨機請假練出千萬支小隊，考試時全員到齊、小聲說話。

給家長和老師：本課講的是 Srivastava、Hinton、Krizhevsky、Sutskever、Salakhutdinov 2014 年 JMLR 第 15 卷的論文 [Dropout: A Simple Way to Prevent Neural Networks from Overfitting](#)。訓練時每個單元以保留概率 p 被保留（隱層常用 0.5，實數輸入層常用 0.8），相當於每次從 2^n 個共享權重的子網絡裡抽一個訓練；測試時不丟棄，把權重乘以 p （權重縮放推斷），近似這些子網絡的平均預測。論文用約 50 個子網絡的蒙特卡洛平均做過對照。實驗覆蓋視覺、語音、文本和可變剪接預測，文本提升最小。

講義中的「音量」是單元輸出乘以 p 的通俗說法，「302 位數」指 2^{1000} 的位數。猜書的數字、「練 100 個模型」的花費、小華的例子、小隊圖、音量圖、估豆子和記詞比賽都是示意，不是原文內容。想深入可看本站的 [第 12 天中文導讀](#)，代碼見開源項目 [30u30 第 12 天目錄](#)（英文）。