

把近路修成直通高速

本課講 2016 年的一篇文章，是第 9 課的續集。何愷明（Kaiming He）和 3 位同事發現：近路上還藏著「關卡」。他們把關卡挪走，讓 1001 層的網絡也能學會。

適合 10 歲左右

約 40 分鐘

需要：鉛筆、紙條、5 位以上同伴

1 熱身：一路打五折

你有 1024 顆糖。每過一個收費站，就要交出一半。把剩下的顆數填進表格。

出發	第 1 站	第 2 站	第 3 站	第 4 站	第 5 站	第 10 站
1024	512					

才過 10 站，糖就只剩 _____ 顆了。如果路上有 1000 個收費站呢？ _____

每一站的規矩都很簡單。可是站數一多，規矩**一站一站疊起來**，結果就面目全非了。今天的故事，就是關於路上的收費站。

2 複習：第 9 課的近路

電腦的「大腦」叫**神經網絡**，由很多層排成一長隊。第 9 課講過，隊伍太長，網絡反而學不好。何愷明他們的辦法是**抄近路**：每幾層旁邊修一條近路，原來的東西沿近路直接送過去。那幾層只學「還差多少」，再補上。

打個比方：老師改作文，不會把作文重寫一遍。她留著原稿，只在旁邊寫批註。原稿就是走近路的東西，批註就是那幾層學到的「還差多少」。

靠這個辦法，網絡一下子長到了一百多層。可是問題沒有完全解決：**1202 層的網絡，反而比 110 層的差**。為什麼？近路不是已經修好了嗎？

3 近路上藏著「關卡」

作者又仔細查了一遍近路。他們發現，原來的近路並不是完全暢通的。近路和小路**合流以後**，還要再過一道「整理站」。

這個整理站會做一件事：把所有小於 0 的數都改成 0。就像溫度計上零下的讀數，全被改成了 0 度。這個步驟有個名字，叫 **ReLU**。每過一塊，就要過一次這樣的關卡。一千層走下來，原來的東西就被改來改去。

作者還做了一組實驗：在近路上故意放各種「東西」，看看會怎樣。

在近路上放了什麼	打個比方	結果
每塊乘 0.5	每站打五折，就像熱身遊戲	變差
加一扇會自己開關的門（試了 2 種裝法）	門有時只開一半	變差
加一個小加工站	每站把東西改造一下	變差
隨機堵住近路	有時候近路不通	變差
什麼都不放	一條直通的路	最好

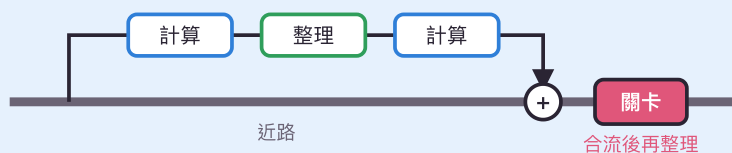
加了「聰明的門」，反而更差？門和加工站每站都會把東西放大或縮小一點。站數一多，就像熱身裡的糖，不是越來越少就是越來越多，網絡就很難學會了。

4 辦法：修一條直通高速路

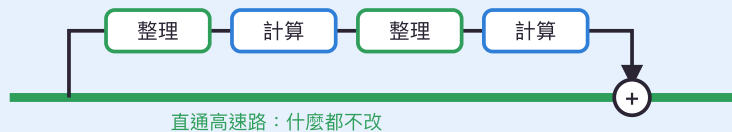
作者的辦法很簡單：**近路上什麼都不放**，合流以後也不再整理。東西在近路上原樣進來，原樣出去，一個字都不改。這樣的近路，就像一條沒有收費站的高速路。

那整理站去哪了？不能扔掉，網絡還需要它。作者把它**挪到旁邊的小路上**，而且放在計算的前面：先整理，再計算。

舊排法（第 9 課）



新排法（第 10 課，叫「預激活」）

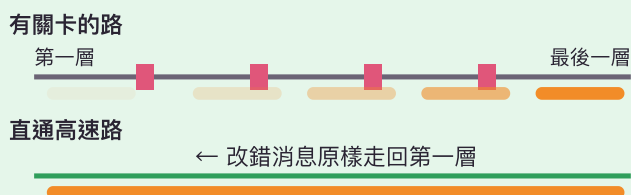


示意圖：每一行是網絡裡的一「塊」。粗線是近路，彎出去的是小路。真正的網絡有幾百塊。

「激活」指的就是整理站裡那一步。原來它在計算後面，叫**後激活**。現在挪到計算前面，就叫**預激活**。「預」就是「預先」的意思。

5 從最後一層到第一層，消息都能走通

第 1 課講過：電腦猜錯了，就悄悄調整自己。每一層怎麼知道該怎麼調？靠一條**改錯消息**。最後一層先看答案對不對，再把「錯在哪裡」一層一層往回傳。每一層聽到消息，就改一改自己。



修了直通高速路，往前送的東西和往回傳的改錯消息，都能原樣走通。最後一層說的話，第一層聽得清清楚楚。

示意圖：橙色條是改錯消息，顏色越淡，消息越模糊。

結果：1001 層也能學會

作者用 **CIFAR-10** 來考網絡。這是一套小圖片，分 10 類：飛機、汽車、鳥、貓、鹿、狗、青蛙、馬、船、卡車。新排法的 **1001 層** 網絡只認錯 4.62%，大約每 100 張錯不到 5 張。用舊排法搭 1001 層，網絡一開始學得特別慢。

6 這和今天有什麼關係

今天的聊天 AI 也是一塊一塊疊起來的，也有一條貫穿到底的主幹道。2019 年的 GPT-2 在文章裡寫明：它把整理步驟放到每一塊前面，和這篇文章的做法類似。後來很多大模型都這樣排：**先整理，再計算，算完加回主幹道，加完不再改。**

這不是唯一的辦法。也有科學家用舊排法配上別的技巧，同樣訓練出了很深的網絡，哪種更好還在討論。

7 動手：高速路傳球

玩法（5 到 9 個人，約 15 分鐘）

- 1 人當出題人，其餘人分成兩隊，各排一列。出題人在兩張紙條上寫同一句口令，含 6 樣東西，例如「紅色 / 3 只 / 小貓 / 週五 / 去公園 / 吃西瓜」。
2. **加工路線（A 隊）**：每人讀紙條 5 秒，翻過去，憑記憶用自己的話寫在新紙條上，只把新紙條傳下去。
3. **直通路線（B 隊）**：原紙條一直往下傳，誰都不準改。每人可以另寫一句補充，貼在旁邊。
4. 傳到最後一人，和原稿對一對。每樣東西沒變畫 ✓，丟了或變了畫 ✕。

路線	1	2	3	4	5	6	保住了幾樣
A 加工路線							
B 直通路線							

你會發現：A 隊的口令站越多變得越多。B 隊的原稿一個字沒變，還多了每個人的補充。這就是直通高速路加小路。

8 動手：排隊報數變形

玩法（6 到 8 個人排成一列，約 10 分鐘，每輪換一條規則）

第 1 輪 打五折站：排頭說 64，後面每人說前一個數的一半。

第 2 輪 放大站：排頭說 2，後面每人說前一個數的 2 倍。

第 3 輪 直通站：排頭說 64，每人接住前一個數，加上抽到的補充卡（-2 到 +2）。

第幾個人	1	2	3	4	5	6	7	8
第 1 輪	64							
第 2 輪	2							
第 3 輪	64							

你會發現：第 1 輪到第 8 人只剩 0.5，第 2 輪卻漲到 256。第 3 輪一直離 64 不遠。

9 詞語小卡片

恆等映射（直走近路）

進來什麼，出去還是什麼，一點都不改。

預激活

小路上先整理、再計算，合流後什麼都不做。

後激活

舊排法：先計算、再整理，合流後還有關卡。

改錯消息（梯度）

從最後一層往回傳，告訴每層怎麼調整。

10 想一想

1. 近路上每塊乘 0.5，看起來只是一點點。為什麼到了幾百層，網絡就學不好了？

2. 近路上裝了「會自己開關的門」，結果反而變差。「東西加得越多越好」對嗎？舉一個生活裡的例子。

3. 這篇文章和第 9 課的出自同一批作者，只隔了 3 個月。你回頭改過自己的作業或作品嗎？改了哪裡？

一句話記住：主幹道上什麼都不改，整理放到小路上先做，1001 層的消息也能從頭傳到尾。

給家長和老師：本課講的是 Kaiming He、Xiangyu Zhang、Shaoqing Ren、Jian Sun 2016 年發表於 ECCV 的論文 [Identity Mappings in Deep Residual Networks](#)（常稱 ResNet v2）。論文分析了恆等捷徑下信號和梯度的直接傳遞，比較了多種捷徑改動和 BN/ReLU 的位置，提出預激活殘差塊（BN → ReLU → 卷積），1001 層網絡在 CIFAR-10 上錯誤率 4.62%。

講義中的簡化與示意：「整理」指 BN 和 ReLU 兩步，「計算」指卷積層；「關卡」指原版殘差塊加法之後的 ReLU；「改錯消息」是對反向傳播梯度的比喻。近路實驗表把論文的 5 種捷徑改動（乘常數 0.5、兩種門控、1×1 卷積、對捷徑做 dropout）合成 4 行。糖果表、兩張示意圖、傳球和報數遊戲都是為本課編的，不是論文內容。想深入可看本站的 [第 10 天中文導讀](#)，代碼見開源項目 [30u30 第 10 天目錄](#)（英文）。