

把近路修成直通高速

本课讲 2016 年的一篇文章，是第 9 课的续集。何恺明（Kaiming He）和 3 位同事发现：近路上还藏着“关卡”。他们把关卡挪走，让 1001 层的网络也能学会。

适合 10 岁左右

约 40 分钟

需要：铅笔、纸条、5 位以上同伴

1 热身：一路打五折

你有 1024 颗糖。每过一个收费站，就要交出一半。把剩下的颗数填进表格。

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

才过 10 站，糖就只剩 _____ 颗了。如果路上有 1000 个收费站呢？ _____

每一站的规矩都很简单。可是站数一多，规矩**一站一站叠起来**，结果就面目全非了。今天的故事，就是关于路上的收费站。

2 复习：第 9 课的近路

电脑的“大脑”叫**神经网络**，由很多层排成一长队。第 9 课讲过，队伍太长，网络反而学不好。何恺明他们的办法是**抄近路**：每几层旁边修一条近路，原来的东西沿近路直接送过去。那几层只学“还差多少”，再补上。

打个比方：老师改作文，不会把作文重写一遍。她留着原稿，只在旁边写批注。原稿就是走近路的东西，批注就是那几层学到的“还差多少”。

靠这个办法，网络一下子长到了一百多层。可是问题没有完全解决：**1202 层的网络，反而比 110 层的差**。为什么？近路不是已经修好了吗？

3 近路上藏着“关卡”

作者又仔细查了一遍近路。他们发现，原来的近路并不是完全畅通的。近路和小路合流以后，还要再过一道“整理站”。

这个整理站会做一件事：把所有小于 0 的数都改成 0。就像温度计上零下的读数，全被改成了 0 度。这个步骤有个名字，叫 ReLU。每过一块，就要过一次这样的关卡。一千层走下来，原来的东西就被改来改去。

作者还做了一组实验：在近路上故意放各种“东西”，看看会怎样。

在近路上放了什么	打个比方	结果
每块乘 0.5	每站打五折，就像热身游戏	变差
加一扇会自己开关的门（试了 2 种装法）	门有时只开一半	变差
加一个小加工站	每站把东西改造一下	变差
随机堵住近路	有时候近路不通	变差
什么都不放	一条直通的路	最好

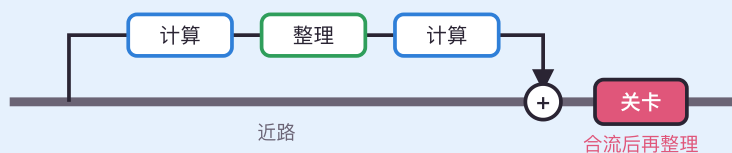
加了“聪明的门”，反而更差？门和加工站每站都会把东西放大或缩小一点。站数一多，就像热身里的糖，不是越来越少就是越来越多，网络就很难学会了。

4 办法：修一条直通高速路

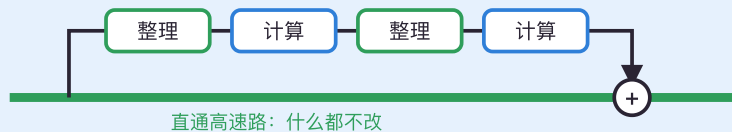
作者的办法很简单：近路上什么都不放，合流以后也不再整理。东西在近路上原样进来，原样出去，一个字都不改。这样的近路，就像一条没有收费站的高速路。

那整理站去哪了？不能扔掉，网络还需要它。作者把它挪到旁边的小路上，而且放在计算的前面：先整理，再计算。

旧排法（第 9 课）



新排法（第 10 课，叫“预激活”）

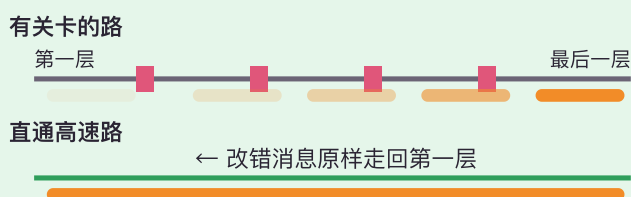


示意图：每一行是网络里的一“块”。粗线是近路，弯出去的是小路。真正的网络有几百块。

“激活”指的就是整理站里那一步。原来它在计算后面，叫**后激活**。现在挪到计算前面，就叫**预激活**。“预”就是“预先”的意思。

5 从最后一层到第一层，消息都能走通

第 1 课讲过：电脑猜错了，就悄悄调整自己。每一层怎么知道该怎么调？靠一条**改错消息**。最后一层先看答案对不对，再把“错在哪里”一层一层往回传。每一层听到消息，就改一改自己。



修了直通高速路，往前送的东西和往回传的改错消息，都能原样走通。最后一层说的话，第一层听得清清楚楚。

示意图：橙色条是改错消息，颜色越淡，消息越模糊。

结果：1001 层也能学会

作者用 **CIFAR-10** 来考网络。这是一套小图片，分 10 类：飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车。新排法的 **1001 层** 网络只认错 4.62%，大约每 100 张错不到 5 张。用旧排法搭 1001 层，网络一开始学得特别慢。

6 这和今天有什么关系

今天的聊天 AI 也是一块一块叠起来的，也有一条贯穿到底的主干道。2019 年的 GPT-2 在文章里写明：它把整理步骤放到每一块前面，和这篇文章的做法类似。后来很多大模型都这样排：**先整理，再计算，算完加回主干道，加完不再改。**

这不是唯一的办法。也有科学家用旧排法配上别的技巧，同样训练出了很深的网络，哪种更好还在讨论。

7 动手：高速路传球

玩法（5 到 9 个人，约 15 分钟）

- 1 人当出题人，其余人分成两队，各排一列。出题人在两张纸条上写同一句口令，含 6 样东西，例如“红色 / 3 只 / 小猫 / 周五 / 去公园 / 吃西瓜”。
- 加工路线（A 队）**：每人读纸条 5 秒，翻过去，凭记忆用自己的话写在新纸条上，只把新纸条传下去。
- 直通过路线（B 队）**：原纸条一直往下传，谁都不准改。每人可以另写一句补充，贴在旁边。
- 传到最后一人，和原稿对一对。每样东西没变画 $\checkmark$ ，丢了或变了画 $\times$ 。

路线	1	2	3	4	5	6	保住了几样
A 加工路线							
B 直通过路线							

你会发现：A 队的口令站越多变得越多。B 队的原稿一个字没变，还多了每个人的补充。这就是直通高速路加小路。

8 动手：排队报数变形

玩法（6 到 8 个人排成一列，约 10 分钟，每轮换一条规则）

- 第 1 轮 打五折站：**排头说 64，后面每人说前一个数的一半。
- 第 2 轮 放大站：**排头说 2，后面每人说前一个数的 2 倍。
- 第 3 轮 直通站：**排头说 64，每人接住前一个数，加上抽到的补充卡（-2 到 +2）。

第几个人	1	2	3	4	5	6	7	8
第 1 轮	64							
第 2 轮	2							
第 3 轮	64							

你会发现：第 1 轮到第 8 人只剩 0.5，第 2 轮却涨到 256。第 3 轮一直离 64 不远。

9 词语小卡片

恒等映射（直通过路）

进来什么，出去还是什么，一点都不改。

预激活

小路上先整理、再计算，合流后什么都不做。

后激活

旧排法：先计算、再整理，合流后还有关卡。

改错消息（梯度）

从最后一层往回传，告诉每层怎么调整。

10 想一想

1. 近路上每块乘 0.5，看起来只是一点点。为什么到了几百层，网络就学不好了？

2. 近路上装了“会自己开关的门”，结果反而变差。“东西加得越多越好”对吗？举一个生活里的例子。

3. 这篇文章和第 9 课的出自同一批作者，只隔了 3 个月。你回头改过自己的作业或作品吗？改了哪里？

一句话记住：主干道上什么都不改，整理放到小路上先做，1001 层的信息也能从头传到尾。

给家长和老师：本课讲的是 Kaiming He、Xiangyu Zhang、Shaoqing Ren、Jian Sun 2016 年发表于 ECCV 的论文 [Identity Mappings in Deep Residual Networks](#)（常称 ResNet v2）。论文分析了恒等捷径下信号和梯度的直接传递，比较了多种捷径改动和 BN/ReLU 的位置，提出预激活残差块（BN → ReLU → 卷积），1001 层网络在 CIFAR-10 上错误率 4.62%。

讲义中的简化与示意：“整理”指 BN 和 ReLU 两步，“计算”指卷积层；“关卡”指原版残差块加法之后的 ReLU；“改错消息”是对反向传播梯度的比喻。近路实验表把论文的 5 种捷径改动（乘常数 0.5、两种门控、1×1 卷积、对捷径做 dropout）合成 4 行。糖果表、两张示意图、传球和报数游戏都是为本课编的，不是论文内容。想深入可看本站的 [第 10 天中文导读](#)，代码见开源项目 [30u30 第 10 天目录](#)（英文）。