

请假请出万千小队

本课讲 2014 年的一篇文章。Srivastava、Hinton、Krizhevsky、Sutskever 和 Salakhutdinov 五位科学家把“请假法”（Dropout）讲了个透：零件随机请假，一个电脑大脑就能顶好多个。

适合 10 岁左右

约 40 分钟

需要：铅笔、一罐豆子、1 个骰子、几位同伴

1 热身：五个人猜书

5 位同学各看了一眼书架，猜有几本书。老师数了一遍：一共 **52 本**。请算出每个人差了多少。

谁猜的	小东	小南	小西	小北	小中
猜了几本	30	70	45	85	20
和 52 差多少					

把 5 个数加起来再除以 5，得到**平均数**：_____ 它和 52 差多少？ _____

谁最准？请圈出来： 小东 小南 小西 小北 小中 平均数

算出来了吗？最准的不是哪一个人，而是**大家的平均数**。这节课的秘密，就藏在这件事里。

2 很多人一起猜，比一个人准

为什么平均数这么准？因为每个人错的方向不一样。有人猜多了，有人猜少了。一平均，多出来的和少掉的就**互相抵消**了一大半。

电脑也是这样。一个练好的电脑大脑，叫一个**模型**。练出 5 个模型，它们死记硬背的地方各不一样。遇到新题，把 5 个模型的答案平均一下，往往比任何一个单独答都准。这个办法叫**集成**，意思是“把好几个模型的意见合在一起”。

可是，集成太贵了

一个大模型要用很多台电脑练好几天。练 100 个，就要花 100 倍的时间。回答一个问题，100 个模型也都要算一遍。有没有办法只练 1 个模型，却得到“很多个模型一起投票”的好处？

3 每次请假，都是一支新小队

第 3 课讲过“请假法”：电脑练习时，每一次都随机选一部分零件请假，剩下的零件照样干活。这篇文章换了一个角度看它：**每一次请假，到场的零件就组成了一支新的小队。**

假设电脑大脑只有 3 个零件：甲、乙、丙。每个零件都有“来”和“不来”两种情况，所以一共有 $2 \times 2 \times 2 = 8$ 种小队：



蓝色实心：到场。灰色虚线：请假。最后一种是全体请假，只是凑数用的。

每多 1 个零件，小队的种数就**翻一倍**。10 个零件有 1024 种小队，20 个零件有 100 多万种。1000 个零件的小队种数，写出来是一个 302 位的数。

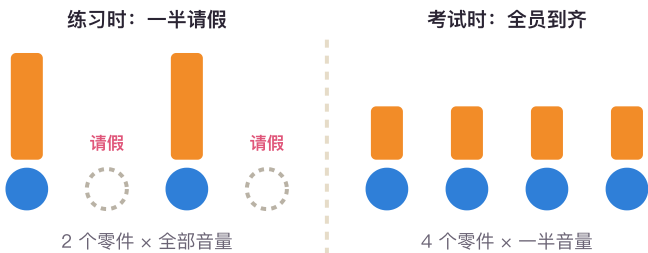
所有小队共用同一套零件

全班每天随机分成不同的小队做任务。小队天天换，同学还是那些同学。小华今天在一支小队学会了算账，明天到了别的小队，本领也跟着带过去。电脑里，大部分小队一次也没轮上练习。可是零件是共用的，它们也跟着变好了。**练 1 个大脑，就等于练了千千万万支小队。**

4 考试时：全员到齐，小声说话

集成要让所有小队投票，再取平均。可是小队多到数不清，一支一支去问，永远问不完。科学家想了一个省事的办法。

考试时谁也不请假，所有零件一起回答。可是每个零件要把音量调小，**按出勤比例打折**。为什么？练习时，每个零件只有一半的时候到场。考试时人全来了，说话声就比平时大了一倍。所以每个零件**只用一半的音量**说话，总音量就和练习时差不多了。



示意图：橙色柱子是每个零件说话的音量。两边加起来的总音量一样。

到场比例不同，折扣也不同。这篇文章里，中间的零件常常一半请假，音量打 5 折。直接看图片的那一排零件大约 8 成到场，音量就打 8 折。

这个省事的办法靠谱吗？

科学家随机抽了大约 50 支小队，分别回答再取平均。结果和“全员到齐、小声说话”差不多。所以考试时只算 1 次，就近似得到了所有小队的平均意见。

5 灵感：爸爸妈妈的基因怎么搭配

科学家说，灵感之一来自生物。你身体里有一套“说明书”，叫基因，它决定你的眼睛颜色、头发卷不卷。你的基因一半来自爸爸，一半来自妈妈。每个孩子拿到的组合都不一样，所以兄弟姐妹长得像，又不完全一样。

一个基因到了下一代，身边的“搭档基因”就换了一批。它没法指望固定搭档，和各种搭档都能合作，才算好基因。

电脑里的零件也会找固定搭档。不请假时，两个零件可能“抱团”：甲只在乙说某句话时才有用，两个合起来死记练习卷上的巧合。这种抱团，科学家叫它共适应。请假法让搭档随时可能不在，抱团就行不通了。就像基因要和各种搭档合作，每个零件也必须自己就有用。

6 看、听、读，样样都管用

这篇文章试得特别多，科学家在 4 类任务上都用了请假法：

看图	认手写数字、认街景照片里的门牌号、认照片里是飞机还是猫
听语音	听一段说话声，分出里面一个一个的音
读文字	读一篇新闻，判断它属于哪一类
生物	预测基因的说明书会怎样被剪开、拼起来

加上请假法以后，几乎每个任务的错误都变少了。其中读文字那个任务进步最小。请假法也有代价。每次只有一部分零件在练，学得慢一些。练习时间大约是原来的 2 到 3 倍。

7 这和今天有什么关系

2012 年，看图程序 AlexNet 用了请假法，在一场看图大赛里遥遥领先。这篇 2014 年的文章把道理讲完整了，请假法从此成了常用工具。后来聊天 AI 用的大脑结构，最初的设计里也用了它。第 3 课说过，最大的聊天 AI 开始大量读书时常常不请假。可是“练习时加一点随机”这个想法，今天很多训练方法都在用。

“很多人一起猜”的道理你也用得上。问聊天 AI 难题时，可以多问几次、多查几处。答案不一致时，就要小心了。

8 动手：估豆子，看看谁最准

玩法（5 人以上，全班更好；约 15 分钟；一个透明罐子装豆子）

- 1. 老师先数好罐子里的豆子，建议 100 到 300 颗，写在纸上藏起来。
- 2. 每人看罐子 20 秒，**不许讨论**，把猜的数偷偷写下来。
- 3. 老师公布真实数量。每人算出自己差了多少，填进下表。
- 4. 把所有人猜的数加起来，再除以人数，得到全班平均数（可以用计算器）。

名字	猜了几颗	和真实数差多少	名字	猜了几颗	和真实数差多少

真实数量：_____ 全班平均：_____ 平均数差了：_____ 比平均数更准的有 _____人
(人多就写在背面)

你会发现：全班平均数常常比大多数人都准，因为猜多的和猜少的抵消了。大家都抄一个人，就没法抵消。请假法的小队也是到场成员不同，犯的错才不同。

9 动手：随机小队记词比赛

玩法（2 队各 3 人加 1 位主持人；约 15 分钟；1 个骰子）

- 1. 每队 3 人编号 1—3。主持人每轮把表里一行词读 2 遍，各队 1 分钟内写下来。
- 2. **固定队**：3 人都在。1 号只记动物，2 号只记水果，3 号只记交通工具。
- 3. **随机队**：练习前掷骰子，1 或 4 是 1 号请假，2 或 5 是 2 号，3 或 6 是 3 号。
- 4. 考试 1 两队全员到齐。考试 2 掷一次骰子，两队同号的人请假。请假的人捂住耳朵。

第几轮	主持人读的 9 个词	固定队	随机队
练习 1	猫、苹果、汽车、狗、香蕉、火车、兔、葡萄、飞机		
练习 2	马、西瓜、轮船、羊、梨、自行车、熊、桃、公交车		
考试 1	鸭、芒果、直升机、虎、樱桃、出租车、蛇、菠萝、马车		
考试 2	鹅、荔枝、火箭、狼、椰子、帆船、鹿、李子、救护车		

右边两列填记住几个词。**你会发现：**考试 2 有人请假时，固定队会整类地丢词。随机队练惯了少一个人，每人都会记好几类，丢得少。

10 词语小卡片

集成

把好几个模型的答案合在一起，比如取平均。通常比一个模型单独答更准，但是很贵。

小队（子网络）

某一次练习时到场的零件。一个大脑里有数不清的小队，共用同一套零件。

全员到齐，小声说话

考试时不请假，每个零件按出勤比例调小音量，近似所有小队的平均意见。

共适应（抱团）

几个零件互相依赖，一起死记练习卷上的巧合。随机请假能拆散这种抱团。

11 想一想

1. 估豆子时，如果大家先讨论、再一起写一个数，平均数还会那么准吗？为什么？

2. 考试时零件全来了，为什么还要“小声说话”？如果大家都用全部音量，会怎样？

3. 你的班级或球队里，有没有“只靠某一个人”的事？他请假了，大家怎样顶上？

一句话记住：随机请假练出千万支小队，考试时全员到齐、小声说话。

给家长和老师：本课讲的是 Srivastava、Hinton、Krizhevsky、Sutskever、Salakhutdinov 2014 年 JMLR 第 15 卷的论文 [Dropout: A Simple Way to Prevent Neural Networks from Overfitting](#)。训练时每个单元以保留概率 p 被保留（隐层常用 0.5，实数输入层常用 0.8），相当于每次从 2^n 个共享权重的子网络里抽一个训练；测试时不丢弃，把权重乘以 p （权重缩放推断），近似这些子网络的平均预测。论文用约 50 个子网络的蒙特卡洛平均做过对照。实验覆盖视觉、语音、文本和可变剪接预测，文本提升最小。

讲义中的“音量”是单元输出乘以 p 的通俗说法，“302 位数”指 2^{1000} 的位数。猜书的数字、“练 100 个模型”的花费、小华的例子、小队图、音量图、估豆子和记词比赛都是示意，不是原文内容。想深入可看本站的 [第 12 天中文导读](#)，代码见开源项目 [30u30 第 12 天目录](#)（英文）。