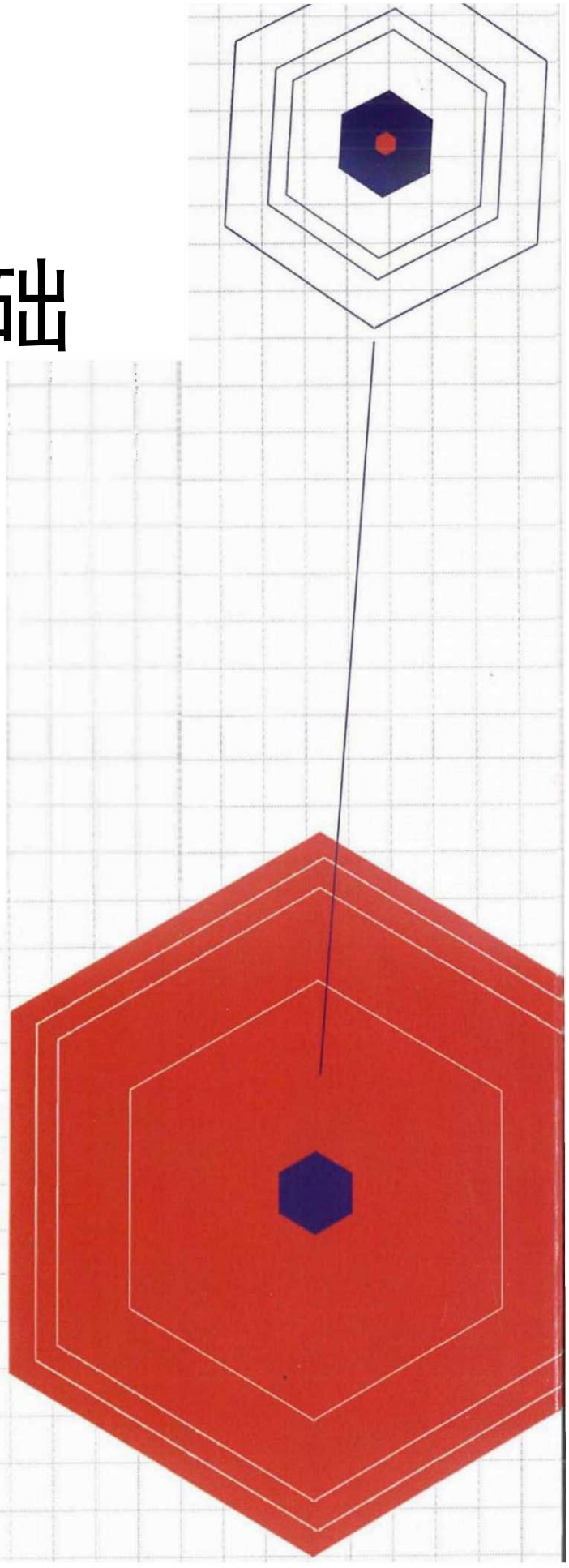
汤晓鸥 陈玉琨 主编

人工智能基础

（高中版）

Fundamentals of Artificial Intelligence



华东师范大学出版社
ECNUP
全国百佳图书出版单位

商務印書館
The Commercial Press
创于1897

人工智能基础

（高中版）

汤晓鸥 陈玉琨 主编

华东师范大学出版社
ECNUP
全国百佳图书出版单位
商務印書館
The Commercial Press

《人工智能基础》编委会

能校园建设，推动人工智能在教学、管理、资源建设等全流程应用"；"广泛开展人工智能科普活动”，“实施全民智能教育项目，在中小学阶段设置人工智能相关课程”；“支持开展人工智能竞赛，开发立体综合教学场、基于大数据智能的在线学习教育平台”。这就对人工智能教育提出了新的任务。

商汤科技是我国人工智能领域的“独角兽”，在该领域有着丰富的积累。2017 年 11 月，商汤与上海市政府签订战略合作协议，目标是使上海成为全球人工智能发展的战略高地。作为有着高度社会良知与责任担当的企业，商汤深知普及人工智能的价值，也在 2017 年 11 月，商汤决定与华东师范大学慕课中心合作，邀请华东师范大学二附中、上海交大附中、上海七宝中学、上海格致中学、 上海市西中学、上海晋元中学教师共同编写人工智能的教材(高中版)。

与你读到过的其他教材不同，本教材以“手脑结合”为主要学习方式，在给你提供必要的基础知识之后，就需要你动手做一些实验，特别希望你发挥独特的想象力，设计在你高中阶段有限的时间中有可能完成的项目，并动手将其转化为独具特色的作品。相信你在这一过程中能够享受到创造的无穷乐趣。

当然，人工智能是一技术含量极高的领域，尤其是一些用到高等数学的算法，可能已经超出了你已学过的范围，有些甚至超出了大学本科的范围，为使你能深入了解人工智能的原理，本教材对这些算法只做定性的介绍，其定量的部分留待你以后再学。商汤已将这些算法开放在其教学实验平台之中，供你在动手时调用。

作为实验教材，错误与不当之处在所难免，欢迎同学们在使用过程中提出批评与改进建议。本教材的目标是：让同学们学会像科学家一样思考。纠错是科学取得进展的基本途径，美国著名心理学家马斯洛(A.H. Maslow)说过：“要做的唯一有气魄的事似乎就是不要害怕错误，投身进去，尽力而为，以期能从大错到纠正它们的过程中学到足够的东西”。①本教材的编者深以为然，更期盼着阅读本教材的同学能以更大的勇气去创新与创造。

编者

2018 年 3 月

① 〔美〕A.H.马斯洛：《自我实现的人》，中译本，三联书店 1987 年版，第 2 页。

目录

第一章 人工智能：新时代的开启

这是一个快速变迁的时代。身处这个时代洪流里的每一个人，无论是课堂中孜孜以求的学子，还是在家中颐养天年的老人，都在享受着日新月 异的便利生活。在这一切便利与舒适的背后，是一场正在深刻地改变着我们的生活与社会的科技浪潮——人工智能。

当你向智能音箱询问今天天气的时候，它通过语音识别技术听懂了你的问题。当你拿起最新的智能手机的时候，它自动解锁打开，因为它通过人脸验证技术认出了主人。当你打开电商网站，你可能在第一时间就看到了自己最喜欢的商品，因为它从你和你的朋友的购买记录中通过大数据分析技术了解了你的兴趣。当你开着新的电动汽车在风景优美的高速公路上疾驰的时候，车载智能系统也在默默地守护着你，时刻不断地监测着可能的危险并适时发出提示。

这一切，仅仅只是一个开始。

很多在十年前仍是科幻小说里的场景，今天已经成为我们真实的生活经历。那么，在人工智能浪潮的驱动下，十年之后我们会生活在什么样的世界里面呢？

1.1 跨越时空：铭铭的一天

一天的开始

2028 年的一个早晨，一缕阳光照进了卧室，铭铭听到了一个柔和的声音：

“铭铭，现在是 2028 年 3 月 29 号早上 7 点，新的一天开始了！”

铭铭对这个声音特别熟悉，它是由智能家居系统所控制的卧室音响发出的。就像一位忠诚的管家，这个系统日复一日分秒不差地照料着铭铭的生活。

铭铭缓缓地从床上坐起来。在他睁开双眼的时候，他看到前面的投影屏被点亮了，屏幕里传来了父亲的问候。

铭铭的父亲是一位著名的人工智能科学家。从小时候开始，铭铭就从父亲的实验室里接触到早期的人工智能技术。有一次，铭铭看到自己的照片在父亲的电脑里被分门别类地做成了拼图。他的活泼，他的欢笑，他的狡黠，在这些拼图中如此真切。父亲告诉他这一切都是电脑通过智能技术自动完成的，这让他感到了莫名的惊喜——这个不经意的瞬间不仅为他打开了一个充满未知的世界，也让他体会到了一直忙于工作的父亲对他的关爱。

铭铭决心把这份父爱传承下去，利用人工智能造福

更多的人。于是，他选择成为一名人工智能工程师。

早餐时间：信息的盛宴

起床后，铭铭来到餐厅。烹饪机器人已经根据铭铭 的口味爱好以及最近几天的健康智能监测系统数据，准备好了一份营养均衡的早餐：一杯奶茶，一盘调配得恰 到好处的沙拉，还有两片他最爱吃的面包。这样一份健康可口的早餐，让他感到精力充沛，心情愉悦。

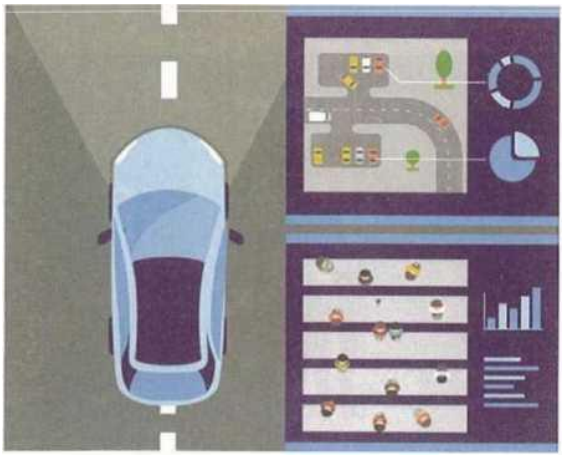
在他吃早餐的时候，餐厅的屏幕开始播放一天的新闻摘要。这是一个信息爆炸的时代，这个城市每天产生 的信息量比起十年前全世界的加起来的还要多。可是铭铭并没有为此而烦恼。一个高效的个性化信息流系统每 天都不断地从海量的新信息中发掘他所关心的部分，并以方便快捷的方式呈现在他面前。

在这个时代，搜索引擎已经不那么重要，新兴的智能网络逐步取代了传统互联网。它们会在合适的时间、合适的地点，以前所未有的效率把信息传递给每一个人。

上班路上：车水马龙间的惬意

从家里出来，铭铭看到他心爱的蓝色电动轿车已经停在了家门前。轿车晚上是在车库里的。智能家居系统一直观察着铭铭的行动，在他出门之前，就提前让车子自主地开到家门前等待着。

铭铭来到车子前面的时候，车门自动打开。上车后，车门又自动关上。在这一系列看似平凡的操作背后是一 个自动身份验证和动作识别模块。在车的主人看来，一切都配合得如此自然。可是当看到一个陌生人靠近时， 车子会保持车门紧闭，并向安全中心发出警报。



在车上，铭铭听到了一个柔和的声音：

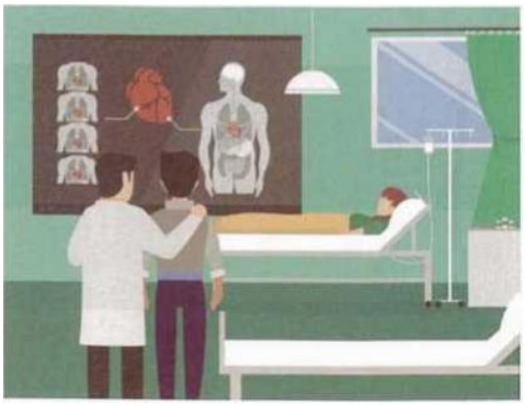
“铭铭，很高兴和您再见面。您现在是要去上班么？”

在铭铭确认后，车子开始启动。因为一个巡回嘉年华刚好来到了这个城市，今天路上的车辆和行人都特别 多。可这并没有给铭铭的座驾带来多大的挑战。在车载激光雷达以及各个方位的视频传感器的帮助下，驾驶系

统准确地检测出了道路上每辆车子、每个行人的方位和动向，精准地调整着行驶的速度和方向。整个车子自如地穿行于车水马龙之间。

参观医院：智能对生命的关怀

今天铭铭约了一个在医院当主治医生的朋友，到他的工作现场实地观看人工智能是如何为医生提供帮助的。在这里，他看到一个危重病例被转介到这里会诊。



病人心脏部位的核磁共振影像显示在大屏幕上。医院里不久前刚升级的智能医疗影像分析系统已经对影像进行了扫描分析，重点区域都被精细标注。这个系统集成了最新的医疗图像模式识别的成果，能够准确地检测数百种不同的病变模式。在这个系统的协助下，医生迅速确定了病因与病情，并和同事们一起设计了两种治疗方案，一种进取，一种保守。助手把两种方案分别输入模拟治疗系统进行模拟测试。根据测试报告.铭铭的朋友和他的同事们很快就判定：使用进取方案是最优的。

从获得影像到确定治疗方案，整个过程只花了几个小时。即使在数年前这也是难以想象的。铭铭的朋友感慨说，这样复杂的病情，在智能医疗系统被引入之前，往往 需要历时多天的会诊，对很多不确定因素进行排查，有时候甚至因此耽误了最佳的治疗时机。智能技术让治疗的效率大大提高了，治疗的效果也有了显著的改善。

下班以后：一次便捷的购物之旅

工作顺利完成，铭铭感到非常愉快。下班后，铭铭决定给自己买一件新的衬衣。于是他驱车来到一家品牌服装专卖店。

和这个城市的很多商店一样，这家老牌专卖店在几年前进行了智能化改造，已经成为一家智能商店。在铭铭进门的时刻，商店的迎宾系统已经把他认出，并在屏幕上显示出对他的欢迎。商店的每一排衣服旁边都立着智能试衣镜，当铭铭拿着衣服来到试衣镜前，镜子中显示出他穿着新衣服时的三维形象。得益于姿态检测与三维人体重建技术的突破，智能试衣镜合成出来的影像非常逼真，而且能自如地配合铭铭摆出的各种姿势，跟真正试穿的体验没有差别。

铭铭对自己新买的衣服非常满意。离开商店后，他第一时间

把试衣的影像发给父亲。在回家的路上，铭铭在座驾的挡风玻璃上看到了父亲的回复：

“嗯，你的品位很有长进，快赶上爸爸了。”

•思考与讨论•

同学们，看完铭铭的故事，对人工智能是不是有了更直接的印象呢？

请谈谈你对未来人工智能生活的想象。

铭铭在十年后的生活是令人憧憬的。这样的生活其实离我们并不遥远。在人工智能浪潮的驱动下，这一切正在一步步地被实现。为了创造新的智能生活，让我们一起来了解人工智能的知识吧。

1.2　光辉岁月：人工智能简史

横空出世

早在上世纪四五十年代，数学家和计算机工程师已经开始探讨用机器模拟智能的可能。

1950 年，艾伦•图灵(Alan Turing)在他的论文《计算机器与智能》(Computing Machinery and Intelligence)中提出了著名的图灵测试(Turing test)。在图灵测试中，一位人类测试员会通过文字与密室里的一台机器和一个人自由对话。如果测试　员无法分辨与之对话的两个实体谁是人谁是机器，则参与对话的机器就被认为通过 测试，虽然图灵测试的科学性受到过质疑，但是它在过去数十年一直被广泛认为是 测试机器智能的重要标准，对人工智能的发展产生了极为深远的影响。

图艾伦·图灵
(1912—1954)

1951 年夏天，当时普林斯顿大学数学系的一位 24 岁的研究生马文•闵斯基 (Manin Minsky)建立了世界上第一个神经网络机器 SNARC (Stochastic Neural Analog Heinforcement Calculator)。在这个只有 40 个神经元的小网络里，人们第一次模拟了神 经信号的传递。这项开创性的工作为人工智能奠定了深远的基础。闵斯基由于他在 人工智能领域的一系列奠基性的贡献，在 1969 年获得计算机科学领域的最高奖图灵奖(Turing Award)。

图 1-2 马文·闵斯基
(1927—2016)

1955 年, 艾伦•纽厄尔(Allen Newell)、赫伯特•西蒙(Herbert Simon)和克里夫•肖(ClifTShaw)建立了一个名为“逻辑理论家”(Logic Theorist)的计算机程序 来模拟人类解决问题的技能-这个程序成功证明了一部大学数学教科书里面 52 个定 理中的 38 个，甚至还找到了比教科书中更优美的证明。这项工作开创了一种日后被 广泛应用的方法：搜索推理(reasoning)。

1956 年，闵斯基、约翰•麦卡锡(John McCarthy)、克劳德•香农(Claude Shannon) 和纳撒尼尔•罗切斯特(Nathan Rochester)在美国的达特茅斯学院组织了一次讨论会。这次会议提出：

“学习和智能的每一个方面都能被精确地描述，使得人们可以制造一台机器来模拟它。”

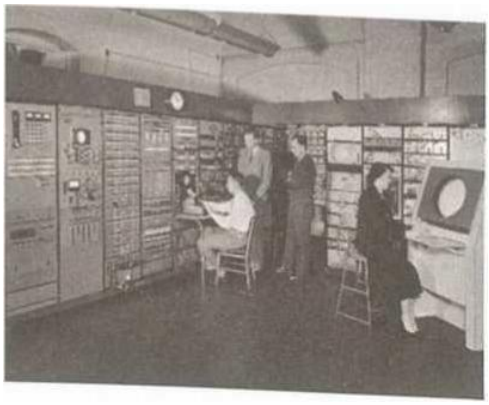
这次会议为这个致力于通过机器来模拟人类智能的新领域定下了名字——“人工智能”(Artificial Inlelligence, AI)，从而正式宣告了人工智能作为一门学科的诞生。

图 1-2 达特茅斯会议，人工智能的诞生

第一次浪潮(1956—1974): 伟大的首航

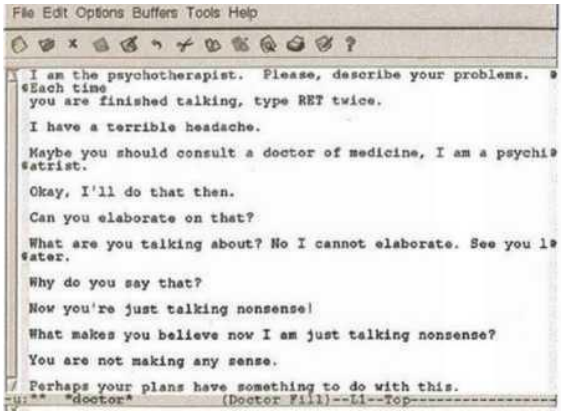
人工智能的诞生震动了全世界，人们第一次看到了智慧通过机器产生的可能。当时有人乐观地预测，一台完全智能的机器将在 20 年内诞生。虽然到现在我们还没看到这样一台机器的身影. 但是它的诞生所点燃的热情确实为这个新生领域的发展注人了无穷的活力。

1963 年，当时刚成立的美国高等研究计划局(ARPA)投入了两百万美元给麻



省理工学院，开启了新项目 Project MAC （ The Project on Mathematics and Computation)。不久后，当时最著名的人工智 能科学家闵斯基和麦卡锡加入了这个项目，并推动了在视觉 和语言理解等领域的一系列研究。Peject MAC 培养了一大批 最早期的计算机科学和人工智能人才，对这些领域的发展产 生了非常深远的影响。这个项目也是现在赫赫有名的麻省理 工学院计算机科学与人工智能实验室（MITCSAIL)的前身。

在巨大的热情和投资的驱动下，一系列新成果在这个时期应运而生。麻省理工学院的约瑟夫•维森鲍姆（Joseph Weizen-baum)教授在 1964 年到 1966 年间建立了世界上第一个自然语言对话程序 ELIZA。ELIZA 通过简单的模式匹配和对话规则与人聊天。虽然从今天的眼光来看这个对话程序显得有点简陋，但是当它第一次展露在世人面前的时候，确实令世人惊叹。日本早稻田大学也在 1967 年到 1972 年间发明了世界上第一个人形机器人，它不仅能对话，还能在视觉系统的引导下在室内走动和抓取物体。



期望越高，失望越大。虽然人工智能领域在诞生之初的成果层出不穷，但还是难以满足社会对这个领域不切实际的期待。由于先驱科学家们的乐观估计一直无法实 现，从 70 年代开始，对人工智能的批评越来越多，在领域内，百花齐放的背后各种 问题也逐步显露出来。一方面，有限的计算能力和快速增长的计算需求之间形成了尖锐的矛盾；另一方面，视觉和自然语言理解中巨大的可变性与模糊性等问题在当时的 条件下构成了难以逾越的障碍。随着公众热情的消退和投资的大幅度削减，人工智能 在 70 年代中期进入了第一个冬天。

第二次浪潮(1980—1987)：专家系统的兴衰

进入 80 年代，由于专家系统(expert system)和人工神经网络(artificial neural network)等技术的新进展，人工智能的浪潮再度兴起。

专家系统是一种基于一组特定规则来回答特定领域问题的程序系统。早在 20 世纪 60 年代，爱德华•费根鲍姆(Edward Feigenbaum)已经开始了对专家系统的早期研究。他因此被称为“专家系统之父”。在 70 年代，斯坦福大学的科学家们开发了一 套名为 MYCIN 的系统，它可以基于 600 条人工编写的规则来诊断血液中的感染。

到了 1980 年，卡耐基梅隆大学为迪吉多公司(DEC)开发了一套名为 XCON 的专家系统，它可以帮助迪吉多公司根据客户需求自动选择计算机部件的组合。这套系 统当时每年可以为迪吉多公司节省 4000 万美元。XC0N 的巨大商业价值极大激发了 工业界对人工智能尤其是专家系统的热情。

值得一提的是，专家系统的成功也逐步改变了人工智能发展的方向。科学家们开始专注于通过智能系统来解决具体领域的实际问题，尽管这和他们建立通用智能的初衷并不完全一致。

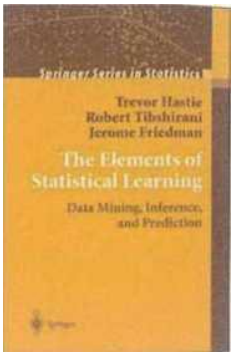
与此同时，人工神经网络的研究也取得了重要进展。1982 年，约翰•霍普菲尔 德(John Hopfield)提出了一种新型的网络形式，即霍普菲尔德神经网络(Hopfield net),在其中引入了相联存储(associative memory)的机制 。1986 年，大卫. 鲁梅尔 哈特(David Rumelhart)、杰弗里•辛顿(Geoffrey Hinlon)和罗纳德•威廉姆斯 (Ronald Williams)联合发表了有里程碑意义的经典论文：《通过误差反向传播学习表 示》(Learning representations by back-propagating errors)。在这篇论文中，他们通过实验展示，反向传播算法(backpmpagation)可以在神经网络的隐藏层中学习到对输入 数据的有效表达。从此，反向传播算法被广泛用于人工神经网络的训练。

图 1-4 杰弗里·辛顿 (1947—)

在新一次人工智能浪潮兴起的同时，日本通商产业省在 1982 年雄心勃勃地开始了旨在建造“第五代计算机”的大型研究计划。这个计划的目标是通过大规模的并行计算来达到类似超级计算机的性能并为未来的人工智能发展提供平台。遗憾的是， 经过了 10 年研发，耗费了 500 亿日元，这个项目未能达成预期的目标。

到了 80 年代后期，产业界对专家系统的巨大投入和过高期望开始显现出负面的效果。人们发现这类系统开发与维护的成本高昂，而商业价值有限。在失望情绪的影响下，对人工智能的投入被大幅度削减，人工智能的发展再度步入冬天。

第三次浪潮(2011 年至今)：厚积薄发，再造辉煌



时间到了 90 年代，历经潮起潮落的人工智能已经进入不惑之年。虽然年轻时气吞天下的雄心遭遇了挫折，但这个领域也变得愈发坚韧。科学家们放下了不切实际的 目标，开始专注于发展能解决具体问题的智能技术。

在这段时期里，研究人工智能的学者开始引入不同学科的数学工具，比如高等代数、概率统计与优化理论，这为人工智能打造了更坚实的数学基础。数学语言的广眨运用，打开了人工智能和其他学科交流合作的渠道，也使得成果能得到更为严谨的检 脸。在数学的驱动下，一大批新的数学模型和算法被发展起来，比如，统计学习理论 (statistical learning theory),支持向量机(support vector machine)、概率图模型(probabilistic graphical model)等。新发展的智能算法被逐步应用于解决实际问题，比如安防 监控、语音识别、网页搜索、购物推荐、与自动化算法交易等等。

新算法在具体场景的成功应用，让科学家们看到了人工智能再度兴起的曙光。

进入了 21 世纪，全球化的加速以及互联网的蓬勃发展带来全球范围电了数据的爆炸性增长。人类迈入了“大数据”时代。与此同时，电脑芯片的计算能力持续高速增长。当前一块 NVIDIA Tesla V100 图形处理器的计算能力已经突破了每秒 10 万亿次浮点运算，超过了 2001 年全球最快的超级计算机。

在数据和计算能力指数式增长的支持下，人工智能算法也取得了重大突破。在 2012 年一次全球范围的图像识别算法竞赛 ILSVRC（也称为 Image Net 挑战赛）中，多伦多大学开发的一个多层神经网络 Alex Net 取得了冠军，并大幅度超越了使用传统 机器学习算法的第二名。这次比赛的成果在人工智能学界引起了广泛的震动。从此, 以多层神经网络为基础的深度学习被推广到多个应用领域，在语音识别、图像分析、 视频理解等诸多领域取得成功。2016 年，谷歌(Google)通过深度学习训练的阿尔法 狗(AlphaGo)程序在一场举世瞩目的比赛中以 4 比 1 战胜了曾经的围棋世界冠军李 世石(shi)。它的改进版更在 2017 年战胜了当时世界排名第一的中国棋手柯洁。

这一系列让世人震惊的成就再一次点燃了全世界对人工智能的热情。世界各国的政府和商业机构都纷纷把人工智能列为未来发展战略的重要部分。由此，人工智能的发展迎来了第三次热潮。

1.3 百花齐放：人工智能在各行各业的应用

近年来，人工智能技术已经被广泛应用于各个行业，并为它们的发展升级注入了新的动力。下面是几个重要的例子。

安防

伴随着城市化的进程和社会经济的高速发展，安全逐步成为全社会共同关心的议题，从平安城市建设到居民社区守护，从公共场所的监控到个人电子设备的保护，我们都离不开一个高效可靠的安全体系。近年来，人工智能技术被大量运用在安防领域，成为我们大家的守护神。

从 2015 年开始，全国多个城市都在加速推进平安城市的建设，积极部署公共安全视频监控体系，希望实现对城市主要道路和重点区域的全覆盖。面对海量的监控视 频，传统的依赖公安民警通过观看视频找出重要片段的方式显然已经不可行了。于是，基于人工智能的视频分析技术被普遍采用。新的智能视频分析技术可以代替民警做很多事情：

. 实时从视频中检测出行人和车辆。

. 自动找到视频中异常的行为（比如，醉酒的行人或者逆行的车辆），并及时发出带有具体地点方位息信的警报。

•自动判断人群的密度和人流的方向，提前发现过密人群带来的潜在危险，帮助工作人员引导和管理人流。

这些技术能把我们的城市管理者从繁重的监控工作中解放出来，更高效地为市民大众服务。

医疗

健康的身体对我们每个人都特别重要。一旦疾病来临，我们就需要及时寻求医生的帮助，解除疾病的痛苦。虽然经过几十年的发展建设，我国的医疗条件比起过去有了很大进步，但是有经验的医生依然是很紧缺的。

人工智能在医疗中的应用为解决“看病难”的问题提供了新的思路。目前，世界各国的诸多研究机构都投入很大的力量开发对医学形像进行自动分析的技术。这此技术可以自动找到医学影像中的重点部位，并进行对比分析。人工智能分析的结果可以为医生诊断提供参考信息，从而有效地减少误诊或者漏诊。除此以外，有些新技术还能通过多张医疗影像重建出人体内器官的三维模型，帮助医生设计手术，确保手术更加精准。

随着智能医疗技术的进步，我们相信人工智能不仅能为医生提供更直接更精准的诊断和治疗建议，而且可以为我们每个人提供健康建议和疾病风险预警，从而让我们得更生活加健康。

智能客服

随着互联网和电子商务的发展，我们和商家的交流变得越来越多元和直接。比如，我们看上一件商品或者一项服务，可以直接通过电话或者网络聊天工具向商家咨询。如何高效地处理来自客户的频繁交流给商家带来了很大的挑战，这也是在互联网时代商家维持竞争力的关键一环。

为了应对这种新的挑战，很多企业开始引入人工智能技术打造智能客服系统。智能客服可以像人一样和客户交流沟通。它可以听懂客户的问题，对问题的意义进行分析（比如客户是询问价格呢还是咨询产品的功能呢），进行准确得体并且个性化的回应，从而提升客户的体验。对企业来说，这样的系统不仅能够提高回应客户的效率，还能自动地对客户的需求和问题进行统计分析，为之后的决策提供依据。

目前，智能客服已经在多个行业领域得到应用，除了电子商务外，还包括金融、通信、物流和旅游等等。

自动驾驶

在现代社会，驾驶汽车上班或者出游成为人们常见的活动。一直以来，开车都是人的专利。随着科技的发展， 人们开始探索让机器自动驾驶的可能。

•2004 年，美国国防部高级研究计划局（DARPA)在莫哈维沙漠(Mojave Desert)举办无人车挑战赛。当时，15 个竞争者无一能完成在无人驾驶的条件下穿越沙漠行驶 142 英里的目标。不过，当时竞争者们为参赛提出的方案 成为当代自动驾驶汽车的雏形。

2010 年，谷歌宣布正在开发自动驾驶汽车，并于一年后在莫哈维沙漠对汽车进行测试。到 2012 年，谷歌宣布其自动驾驶汽车已经行驶了 30 万英里，并且没有发生过事故。

-2014 年，百度和宝马（BMW)宣布开始自动驾驶研究。

到目前为止，自动驾驶研究的大幕已经拉开，有多家公司投入到了自动驾驶技术的研发当中。

现在的自动驾驶汽车通过多种传感器，包括视频摄像头、激光雷达、卫星定位系统（北斗卫星导航系统 BDS、 全球定位系统 GPS 等）等，来对行驶环境进行实时感知。智能驾驶系统可以对多种感知信号进行综合分析，通过结

合地图和指示标志(比如交通灯和路牌)，实时规划驾驶路线，并发出指令，控制车子的运行。

工业制造

我国是工业大国.随着各种产品的快速迭代以及现代人对于定制化产品的强烈需求，工业制造系统必须变得更加“聪明”，而人工智能则是提升工业制造系统的最强动力。

比如，品质监控是生产过程中的重要环节，一个质量不过关的零件如果流向市场，不仅会使消费体验大打折扣，更有可能导致严重的安全事故。因此传统生产线上都安排大量的检测工人用肉眼进行质量检测。这种人工检测方式 不仅容易漏检和误判，更会给质检工人造成疲劳伤害。因此很多工业产品公司开发使用人工智能的视觉工具，帮助 工厂自动检测出形态各异的缺陷。

2011 年汉诺威工业博览会(Hannover Messe)上，德国提出了工业 4. 0 概念，其中最重要的就是在工业环境中使 用大母的传感器采集海量的数据。人工智能则成为分析这些海量数据并从中挖掘有价值信息的强大武器。西门子 (Siemens)和通用电气(GE)等工业巨头纷纷开发了人工智能系统，用来预测生产环节的风险、降低材料浪费和能源损耗，并同时提升生产效率。

1.4 初露真容：人工智能与机器学习

什么是人工智能

对干人工智能的定义，学界一直有不同的表述。在这里，我们采用一种被广泛接受的说法:

人工智能是通过机器来模拟人类认知能力的技术。

人工智能涉及很广，涵盖了感知、学习、推理与决策等方面的能力。从实际应用的角度说，人工智能最核心的能力就是根据给定的输入做出判断或预测， 比如：

•在人脸识别应用中，它是根据输入的照片，判断照片中的人是谁。

•在语音识别中，它可以根据人说话的音频信号，判断说话的内容。

•在医疗诊断中，它可以根据输入小医疗彩像，判断疾病的成因和性质。

•在电于商务网站中，它可以根据一个用户过去的购买记录，预测这位用户对什 么商品感兴趣，从而让网站做出相应的推荐。

•在金融应用中，它可以根据一只股票过去的价格和交易信息，预测它未来的价格走势。

•思考和讨论•

你在日常生活中接触到的人工智能技术，它们是根据什么输入，做出什么样的预测和判断呢？

那么人工智能是如何自动做出判断或预测的呢？其实这并不神秘，有时候我们仅需要一些简单的规则。比如，我们用生活中常见的体温计就可以组成一个非常简单的智能系统。它通过水银或者其他对温度敏感的物质获得体温读数作为输入，然后通过简单的规则，比如“体温是否超过 37.5 摄氏度”，来判断接受测量的人是否正在发烧。

在 80 年代一度兴起的专家系统就是基于人工定义的规则来回答特定问题的。可是人工定义规则的方式有着很多局限性。一方面，在复杂的应用场景下建立完备的规则系统往往是一个非常昂贵而耗时的过程；另一方面，很多基于自然输入的应用，比如语音和图像的识别，很难以人工的方式定义具体的规则。因此，当代的人工智能普遍通过学习(learning)来获得进行预测和判断的能力。这样的方法被称为机器学习(machine learning)，它已经成为人工智能的主流方法。

从数据中学习

机器学习方法通常是从已知数据(data)中去学习数据中蕴含的规律或者判断规则。但是，已知数据主要是用作学习的素材，而学习的主要目的是推广(generalize)，也就是把学到的规则应用到未来的新数据上并做出判断或者预测。

机器学习有多种不同的方式。最常见的一种机器学习方式是监督学习(supervised learning)。下面我们看一个例子。这里，我们希望能得到一个公式来预测一种宝 石的价格。而我们知道这种宝石的价格主要由它的重量和等级确定。如果我们使用监 督学习的方法，为了得到这个价格公式，我们需要先收集一批宝石价格的数据，如表 1-1。

表 1-1 宝石的价格

重量	等级	价格
3	2	7030
4	1	6010
2	3	7960

现在我们准备根据表 1-1 来学习一个可用于价格预测的公式。表中每一行称为一个样本（sample)。我们可以看到，每个样本包含了两个部分：用于预测的输入信息（重量、等级）和预测量（价格）的真实值。通过表 1-1，我们可以对不同的预 测进行公式测试，并通过比较在每个样本上的预测值和真实价格的差别获得反馈。机 器学习的算法然后依据这些反馈不断地对预测的公式进行调整。在这种学习方式中，预测量的真实值通过提供反馈对学习过程起到了监督的作用。我们称这样的学习方式 为监督学习。在实际应用中，监督学习是一种非常高效的学习方式。我们会在后面的章节中介绍监督学习的具体方法。

监督学习要求为每个样本提供预测量的真实值，这在有些应用场合是有困难的。 比如在医疗诊断的应用中，如果要通过监督学习来获得诊断模型，则需要请专业的医 生对大量的病例及它们的医疗影像资料进行精确标注。这需要耗费大量的人力，代价 非常高昂。为了克服这样的困难，研究者们也在积极探索不同的方法，希望可以在不 提供监督信息（预测量的真实值）的条件下进行学习。我们称这样的方法为无监督学习（unsupervised learning)。无监督学习往往比监督学习困难得多，但是由于它能 帮助我们克服在很多实际应用中获取监督数据的困难，因此一直是人工智能发展的一

个重要研究方向。

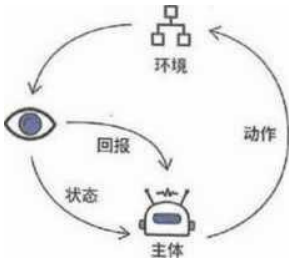
近年来，还有另外一种被称为半监督学习的学习方式也受到了广泛关注。半监督学习(semi-supervised learning)介于监督学习与无监督学习之间，它要求对小部分的样本提 供预测虽的真实值。这种方法通过有效利用所提供的小部分监督信息，往往可以取得比无监督学习更好的效果，同时也把获取监督信息的成本控制在可以接受的范围。

在行动中学习

在机器学习的实际应用中，我们还会遇到另一种类型的问题：利用学习得到的模型来指导行动。比如在下棋、股票交易或商业决策等场景中，我们关注的不是某个判断是否准确，而是行动过程能否带来最大的收益。为了解决这类问题，人们提出了一 种不同的机器学习方式，称为强化学习（reinforcemenl learning)。

强化学习的目标是要获得一个策略（policy)去指导行动。比如在围棋博弈中，这个策略可以根据盘面形势指导每一步应该在哪里落子；在股票交易中，这个策略会 告诉我们在什么时候买入、什么时候卖出。与监督学习不同，强化学习不需要一系列包含输入与预测的样本，它是在行动中学习。

一个强化学习模型一般包含如下几个部分：



•一组可以动态变化的状态（state)。比如，围棋棋盘上黑白子的分布位置，市场上每只股票的价格。

•一组可以选取的动作（action)。比如，对于围棋来说，就是可以落子的位置；对于股票交易来说，就是每个时间点，买入或者卖出的股票以及数量。

•一个可以和决策主体（agent）进行交互的环境（environment)。这个环境会决定每个动作后状态如何变化。比如，围棋博弈中的对手，或者股票市场。在强化学习中，为了降低学习的代价，很多时候我们会用一个通过机器模拟的环境，而不是以真实场景作为环境。

•回报(rewand)规则。当决策主体通过行动使状态发生变化时，它会获得回报或者受到惩罚（回报为负值）。

强化学习会从一个初始的策略开始。通常情况下，初始策略不一定很理想。在学习过程中，决策主体通过行动和环境进行交互，不断获得反馈（回报或者惩罚），并根据反馈调整优化策略。这是一种非常强大的学习方式。持续不断的强化学习甚至获 得比人类更优的决策机制。在 2016 年击败围棋世界冠军李世石九段的阿尔法狗，其令世人震惊的博弈能力就是通过强化学习训练出来的。

1.5 本章小结

人工智能是研究如何通过机器来模拟人类认知能力的学科。它可以通过人工定义或者从数和据行动中学习的方式获得预测和决策的能力。通过过去几十年的努力，人 工智能已经获得了长足的发展，并且在多个行业得到了成功的应用。

人工智能这一新兴的科技浪潮正在深刻地改变看我们的世界并影响着我们的生活，但是这一切仅仅只是一个开始。我们的生产、生活、社交、娱乐等方方面面依然 可以通过人工智能技术的应用得到进一步的提升。人工智能过去的发展为我们展现了 一个令人激动的前景，而这个更美好的新时代需要我们共同努力去创造。

第二章 牛刀小试：察异辨花

铭铭去郊游，偶到一处突然被深深吸引：好 一片山花烂漫！他再也不想多走一步，缠着爸爸问道：“这是什么花呀？”“这叫鸢(yuan)尾。”细心的铭铭发现在角落处有几株花瓣略有不同，“这也是鸢尾花吗？”“这是另一种鸢尾，它们的 花瓣大小不同，你可以用手比画一下，记录下它们的……”爸爸的话音未落，铭铭就跑开指向远处：“爸爸，我们还是去那边看看吧！”

看到一张图片，我们能够分辨图片上有什么动物，是猫还是狗；听到一首歌曲，我们能够区分是古典音乐还是流行音乐；看到一段视频，我们知道里面的演员是在舞蹈还是在长跑……在生活中，我们经常会判断一个事物的类型，这样的过程在人工智能领域里被称为分类。

2.1 初学乍练：分类任务

人工智能系统处理的是各种各样的数据：图像、声音、文字、视频等等。图 2-1 中展示了常见的数据类型和与它们相关的一些应用。数据(data)是信息的载体。分类(classification)就是要根据所给数据的不同特点，判断它属于哪个类别。

在这一章，我们学习一个简单的分类任务——对鸢尾花(iris)的两个品种进行分类。鸢尾花的花瓣鲜艳美丽，叶片青翠碧绿，让人赏心悦目。全世界大约有 300 个品种，常见的有变色鸢尾*(Iris wrskobr)*和山鸢尾*(Iris setose)*。如图 2-2 所示，它们有着形状与色彩相似的花瓣和萼片。一般来说，变色鸢尾有较大的花瓣.而山鸢尾的花瓣较小。我们通过对鸢尾花的分类这个例子来了解分类问题中的基本概念和流程。

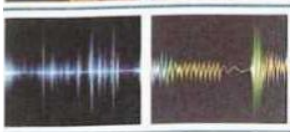
数据类型	示意图	相关应用举例
图像（照片、图画） 详见第三章		· 基于图像的人脸识别、物体分类、场景理解 · 图像处理、艺术风格转换、 图片生成
声音（语音、音乐） 详见第四章		· 语音识别、机黠对话、对话理解 · 听歌识曲、自动谱曲
视频（影视） 详见第五章		· 视频分类、内容理解、查找检索 · 视频总结、自动生成视频
文字 详见第七		· 文章聚类、阅读理解 · 自动生成文章
其他数据类型 详见第九章等		· 棋谱数据在 AlphaGo 中使用 · 基因测序数据辅助疾病的预测诊断

图 2-1 常见的数据类型及 其应用

变色鸢尾

山鸢尾

图 2-2 变色爲尾和山莺尾

我们想要构建一个简单的人工智能系统，它能够像人类一样区分变色鸢尾和山鸢尾。像这样完成分类任务的人工智能系统，被称为分类器(classifier)。图 2-3 展示了整个系统的流程。当看到一朵鸢尾花时，首先提取它的特征，然后将这些特征输人 到训练好的分类器中，分类器就能够根据这些特征做出预测，输出鸢尾花的品种。在 接下来的小节中，让我们一步一步地构建出这个系统吧。

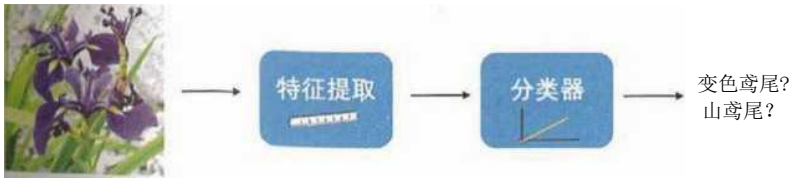


图 2-3 区分鸢尾花品种的人工 智能系统

2.2 含英咀华：提取特征

我们往往会根据物体具有的一些特点来区分它们，比如辨别不同鸢尾花品种的时候，依据的是鸢尾花的花瓣大小。像这种可以对事物的某些方面的特点进行刻画的数字或者属性，我们称为特征(feature)。

在鸢尾花分类中，怎么得到可以被人工智能系统所使用的特征呢？经过尝试，人们发现用花瓣的长度和宽度作为鸢尾花的特征，可以让分类器有效地分类。提取特征时，如图 2-4 左图所示，直接用尺子测量即可，选用这样的 特征也符合人们根据鸢尾花花瓣大小来区分种类的生活经验。

图 2-4 提取鸢尾花的不同特征

特征是在分类器乃至于所有人工智能系统中非常重要的概念。对同样的事物，我们可以提取出各种各样的特征。 如图 2-4 右图所示，我们也可以用鸢尾花植株的高度或者花瓣颜色作为特征。但是，鸢尾花的植株高度和品种没有 直接关系，一朵鸢尾花在生命的不同阶段也有着不同的高度；不同鸢尾花品种又都有着颜色相近的花瓣，所以用鸢尾花的植株高度和花瓣颜色很难有效区分鸢尾花的品种。我们看到，不同的特征对于分类器的准确分类会有很大的影响。

因此，我们需要根据物体和数据本身具有的特点，考虑不同类别之间的差异，并在此基础上设计出有效的特征。 而这不是一件简单的事一它往往需要我们真正理解事物的特点和不同类型之间的差异。特征的质量很大程度上决 定了分类器最终分类效果的好坏。

上面例子中使用的花瓣长度和宽度是较为简单的特征。在后面章节中，我们会针对不同类型的数据，逐步介绍几种常用的人工设计的特征。比如，对于图像，人们设计出了方向梯度直方图；对于声音，人们设计出了梅尔频率倒谱系数；对于视频，有光流直方图；对于文本，有词频率-逆文档频率等。

特征向量

通过实际的测量，我们得到了鸢尾花的特征——花瓣的长度和宽度，那么在数学上它们如何表达呢？我们可以 用 x_1来表示花瓣的长度，用 x_2来表示花瓣的宽度。为了使用方便，进一步地把这两个数字一起放进括号中，写成 (x_1, x_2)这种形式的一组数据在数学中被称为向量(vector)。

知识链接：向量和向量运算

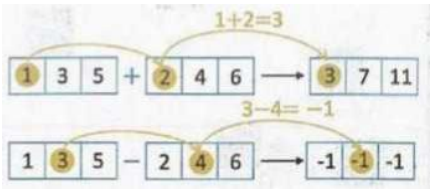
在数学上，向量就是多个数字按序排成一行，比如(1, 3, 5)。其中数字的个数称为向量的维数(dimen- sion)。例如,(1, 3, 5)的维数是 3，我们说，这是一个三维向量。

向量还可以进行简单的运算。

加减法：两个相同维数的向量相加减，就是它们的每个数字对应相加减。

(1, 3, 5) + (2, 4, 6) = (1 + 2, 3 + 4, 5 + 6) = (3, 7, 11)

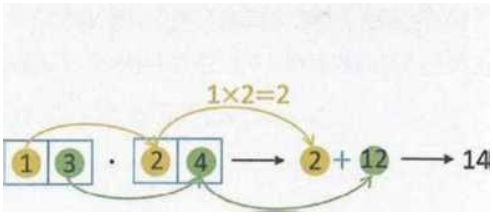
(1, 3, 5) - (2, 4, 6) = (1 - 2, 3 - 4, 5 - 6) = (-1, -1, -1)



数量乘法：一个数和向量相乘，就是这个数和向量中的每一个数字相乘。

5 x (1, 3, 5) = (5x1, 5 x3, 5 x5) = (5, 15, 25)

内积：两个具有相同维数的向量做内积，就是它们的每个数学对应相乘并求和。



有了向量这个数学工具后，我们就可以把描述一个事物的特征数值都组织在一起，形成一个特征向量(feature vector),对它进行更完备的刻画。一般地，一个 n 维 的特征向量可以被表示为 x=(x_1, x_2....., x_n)。比如如测量得到一朵鸢尾花的花瓣长度为 1.1 厘米，宽度厘米为 0.1 厘米，那么这朵鸢尾花的特征就可以用(1. 1, 0. 1)来表示。

特征点和特征空间

有了特征的向最表示之后，进一步地，我们可以把特征向量表示在直角坐标系中。比如(1. 1, 0. 1),就可以看成是直角坐标系中的一个点。

如图 2-5 所示，我们把鸢尾花的特征向量画在了坐标系中。坐标系中的一个点就代表了一朵鸢尾花的特征，这些表示特征向量的点被称为特征点(feature point); 所有这些特征点构成的空间被称为特征空间(feature space)。

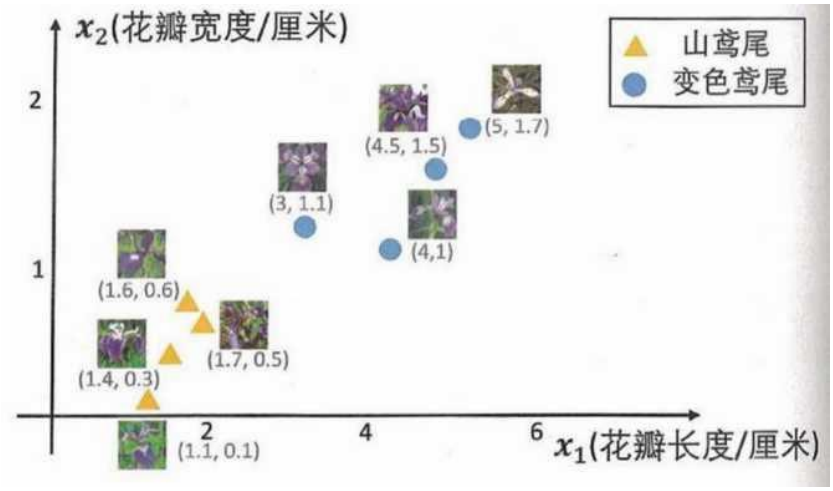


图 2-5 特征向量在直角坐标系中的表示

在图 2-5 所示的特征空间中，特征点到特征点之间的平面距离可以用来衡量鸢尾之间的相似程度。一般地. 对于任意维数的特征空间，我们都可以使用特征点之间的距离(distance)来衡量物体之间的相似程度。高维特征空间的距离计算公式与二维类似，比如在三维空间里，有两个点分别表示为(x_1，x_2，x_3)和(z_1，z_2, z_3),那么这两个点之间的距离 d 可以通过下面的式子进行计算：

$$d = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + (x_3 - z_3)^2}$$

2.3 分门别类：分类器

分类器就是一个由特征向量到预测类别的函数。在鸢尾花的分类问题中，我们用 +1 和-1 两个个值分别代表变色鸢尾和山鸢尾两个类别，并用字母 y 表示，即 y 可以 取+1 和-1 两个值。前面我们已经提取了鸢尾花的特征，将它表示为特征向量，并 把特征向量画在了特征空间中。从图 2-6 上看，对鸢尾花品种分类的问题就转变成 在待征空间中将一些特征点分开的问题。如果我们用直线作为分界线，那么这个问题 就变成：坐标平面中有两类点，画一条直线将这两类点分开。

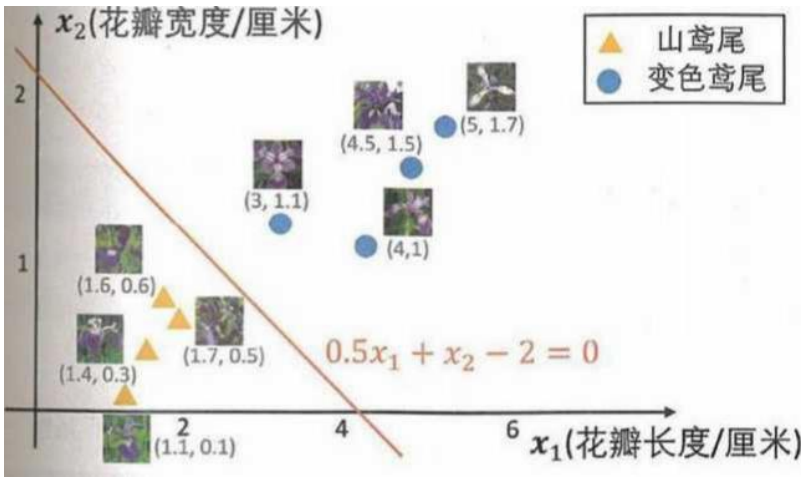


图 2-6 画一条直线来区分变色 鸢尾和山鸢尾

我们可以轻而易举地在图 2-6 中画一条直线 $0.5x_1+x_2-2=0$,它将整个 坐标平面分为两个区域。若使落在直线右上区域的特征点输出+1,代表变色鸢 尾；落在直线左下区域的特征点输出-1,代表山鸢尾，应用这样的规则，我们 就能够得到将鸢尾花正确分类的分类器。这个规则代表的分类器可以用下面的函 数来表示：

$$g(x_1, x_2) = \begin{cases} +1, & 0.5x_1 + x_2 - 2 \geqslant 0 \\ -1, & 0.5x_1 + x_2 - 2 < 0 \end{cases}$$

其中，$0.5x_1+x_2-2$ 和图中所画的直线有着对应关系，我们把它记为 $f(x_1, x_2)$。 如果 $f(x_1, x_2) \geqslant 0$，就表示特征点(x_1, x_2)在直线的右上区域；反之，表示特征 点在直线的左下区域。

f（x）是分类函数 g(x)的核心。f(x）的不同，相当于在图 2-6 中画了不同的线

用来分开不同的类。函数 f(x)的形式多种多样，具有 $f(x_1, x_2, \cdots, x_n) = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b$ 形式的分类器被称为线性分类器（linear classifier),其中 n 是特征向量的维数。a_1，a_2，$\ldots a_n$，b 函数的系数，被称为分类器的参数（parameters)。在上面的例子中，0.5, 1, -2 就是分类器的参数的取值。

图 2-7 分类器是一个由特征向

在区分鸢尾花品种的简单例子中，我们可以直接画出一条直线将两类点分开。实际情况中，特征点在特征空间中的位置分布非常复杂，采用观察和尝试来画出分类直 线往往是不可能的，也是没有效率的。因此，需要通过一些方法，让分类器自己学习 得到分类直线。

训练分类器

我们可以把人工智能系统和人类做类比。人们需要经过在学校的学习来吸收知识；为了检验学习效果，要参加考试；学到知识掌握技能后，就会在工作中解决实际 的问题。人工智能系统也类似，它的学习过程被称为训练（training)；考试过程被称 为测试(testing)；它解决实际问题的过程，被称为应用(application)。

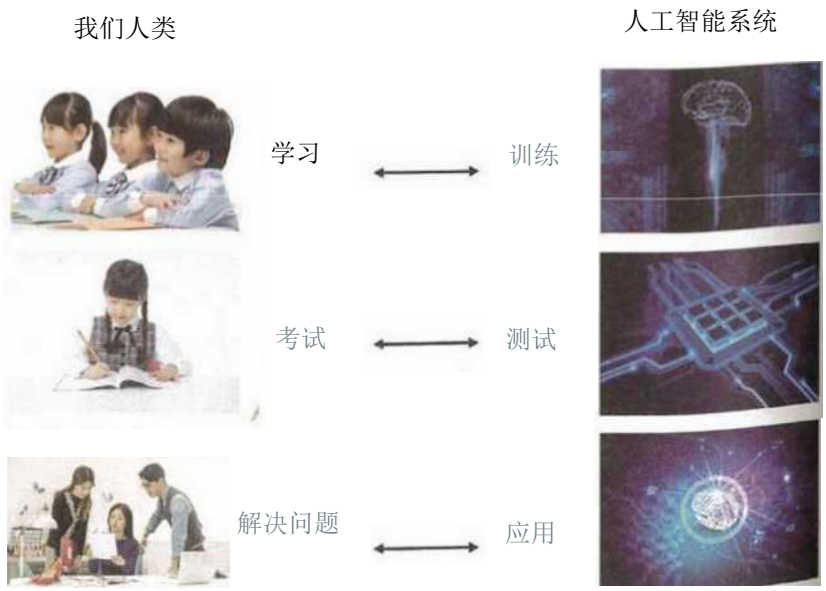


图 2-8　人类和人工智能系统的类比

让分类器学目得到合适参数的过程称为分类器的训练。在本章中，训练分类器就是找到一条好的分类直线。

我们在学校通过老师、课本来接受知识，那么人工智能系统是通过什么进行学习的呢？答案是数据。数据是人工智能的支柱之一，人工智能系统的训练往往需要大量 的数据做支撑。

在训练阶段使用的数据被称为训练数据；相应地，测试阶段使用的数据被称为测试数据。在分类中，训练和测试数据一般都需要知道它们实际的类别。人工地给数据标上真实类别（其他任务中对应着其他真实的值）的过程被称为数据的标注（anno- tation）。数据标注的过程耗时耗力，有的数据标注还可能需要相关领域的专业知识。对数据的采集和标注是非常重要的过程。人工智能系统是数据驱动的。数据标注的质 量会直接影响到训练后人工智能系统的性能好坏。

鸢尾花的数据是由美国植物学家埃德加•安德森（Edgar Anderson)在加拿大加斯佩半岛（Gaspesie)的农场测量鸢尾花的花瓣长度和宽度等特征得到的。此外，他根据自己所学的植物学知识，标注好每一朵花属于什么种类。

鸢尾花	花瓣长度/厘米	花瓣宽度/厘米	类别
	1.1	0.1	山鸢尾
	1.7	0.5	山鸢尾
	1.4	0.3	山鸢尾
	1.6	0.6	山鸢尾
	5.0	1.7	变色鸢尾
	4.0	1.0	变色鸢尾
	4.5	1.5	变色鸢尾
	3.0	1.1	变色鸢尾
.....	.....	.....	.....

图 2-9 埃德加·安德森采集和标注的部分鸢尾花数据

经过埃德加. 安德森的采集和标注后，我们得到了如图2-9所示的数据集。在这个数据集中，每一行代表一个样本，它包含了一朵鸢尾花的特征，以及它对应的类别。有了这样的数据集. 我们就可以在它基础之上去训练一个分类器。当一个数据集被用于分类器训练，我们会称之为训练集(training set)。接下来，我们会介绍基于数

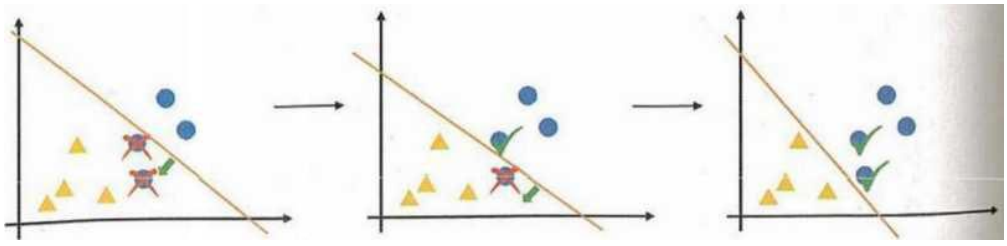
据集来训练分类器的过程。这个过程是由一系列判断和计算的步骤组成的，通常被称为算法(algorithm)。在一个数据集上，使用不同的算法可能会获得不同的分类器。 如何设计一个算法能获得性能好（即分类准确率高）的分类器是机器学习里面一个经典的研究课题。

我们继续上面的例一寻找一个线性分类器对鸢尾花分类。在这里线性分类器中的 f(x)可以被概括地表示为 $f(x_1, x_2) = a_1x_1 + a_2x_2 + b$。的目的就是找到合适的参参数 a_1，a_2，b，使得对应的分类器能够区分变色鸢尾和山鸢尾。

下面我们介绍两种常见的训练线性分类器的算法一感知器和支持向量机，它提供了两种利用训练数据自动寻找参数的方法。

感知器

感知器(perceptron)是一种训练线性分类器的算法，它的主要想法是利用被误分类的训练数据调整现有分类器的参数，使得调整后的分类器判断得更加准确。我们 在图 2-10 中通过简单的示意图来进行说明：最开始分类直线分错了两个样本，分类的直线便向该误分类样本一侧移动；第一次调整以后，一个误分类样本的预测被纠正，但仍有一个样本被误分类一这个仍被误分类的样本到分类直线的距离相比调整之前减小了。接下去，直线向着这个仍被误分类的样本一侧移动，直到分类直线越过，该误分类样本。这样，所有训练数据都被正确分类了。



两个样本被分错，直线向误分类样本一侧移动

一个误分类样本被纠正，直 线向另一误分类样本侧移动

直到所有训练数据都被正确分类

图 2-10 感知器的训练过程示意

感知器学习算法：感知器使用被分错的样本来调整分类器一一如果标注的类别是 +1，使得 $a_1x_1 + a_2x_2 + b < 0$ 的样本就是被误分类；如果标注的类别是-1，使得 $a_1x_1 + a_2x_2 + b \geq 0$ 的样本就是被误分类。我们可以把这两种情况综合起来-----若 $y(a_1x_1 + a_2x_2 + b) \leq 0$，那么样本就被分错了，其中 y 表示数据的真实类别。感知器学

习算法根据这个误分类的样本来调整分类直线的参数，使得直线向该误分类数据一侧 移动，以减小该误分类数据与直线的距离，直到直线越过该误分类数据使其被正确 分类。

具体的感知器学习算法如下所示。

感知器学习算法

第一步：选取初始分类器参数 a_1, a_2 , b；

第二步：在训练集中选取一个训练数据，如果这个训练数据被误分类，即 $y \times (a_1x_1 + a_2x_2 + b) \leq 0$, 则按照以下规则更新参数（将箭头右边更新后的值赋给左边的参数）：

乘号

$$a_1 \leftarrow a_1 + \eta y x_1$$

$$a_2 \leftarrow a_2 + \eta y x_2$$

$$b \leftarrow b + \eta y$$

第三步：回到第二步，直到训练数据中没有误分类数据为止。

其中，η 是学习率（learning rate),学习率是指每一次更新参数的程度大小。

•思考与讨论•

为什么上面的感知器学习算法能够获得一个性能良好的分类器呢。

感知器的学习算法就是不断减少对数据误分类的过程。在这里，同学们可能会有两个疑问：一是如何衡量分类器对数据的误分类程度呢？二是我们该如何利用误分类的数据来调整分类器的参数，也就是感知器学习算法中更新参数的规则是怎么来的呢？为此，我们分别介绍损失函数和优化方法来回答上面的两个问题。

损失函数(loss function)是在训练过程中用来度量分类器输出错误程度的数学化表示。预测错误程度越大，损失函数的取值就越大。定义合适的损失函数对于训练分类器是非常重要的。感知器和支持向量机就是基于不同的损失函数建立起来的。

在分类鸢尾花的例子中，假定总共有 N 个训练数据，我们用 $(x_1^{(i)}, x_2^{(i)})$ 来表示第 i 个训练数据的特征向量，$y^{(i)}$ 表示第 i 个训练练数据的标注类别，那么感知器的损失函数 L 则定义为：

$$L(a_1, a_2, b) = \sum^{N} \max(0, -y^{(i)} \times (a_1 x_1^{(i)} + a_2 x_2^{(i)} + b))$$

知识链接：求和符号和求最大符号

在数学上，我们采用 来方便地表示求和。举个例子，我们想写一个式子表示从 1 到 100 求和。一项项地写 出来显然太麻烦，用求和符号 就可以非常方便地表示：

$$\sum_{i=1}^{100} i = 1 + 2 + \ldots + 99 + 100 = 5050$$

∑符号下面的 $i = 1$ 表示计数的变量是 i，并且从 1 开始；上面的 100 表示计数一直到 $i = 100$ 为止。 我们用 max 符号来表示取最大的运算过程。比如 max(0, -1) = 0, max(0, 1, 2) =2。

上述感知器的损失函数表示对训练数据中每一个样本计算-y x $(a_1x_1 + a_2x_2 + b)$，并和零比-----如果大于零，则损失函数增加；否则损失函数就不变。在感知器学习算法中我们已经知道，使得-y x $(a_1x_1 + a_2x_2 + b) \geq 0$ 的样本是被误分类的样本。因此上面的损失函数相当于定义在所有被误分类的数据上。

显然，如果没有误分类的数据，那么损失函数为零，如果有误分类数据，就会使得损失函数增大；并且误分类数据越多，损失函数越大。

我们利用上面损失函数的公式动手算一算。在图 2-11 (a)中，有一个误分类数据，损失函数的值为 0.2；在图 2-11 (b)中，更多的数据被误分类，损失函数的值增大到 2.65。此外，图 2-11 (b)还告诉我们，在直线确定的情况下，误分类的数据点离直线越远，损失函数越大。

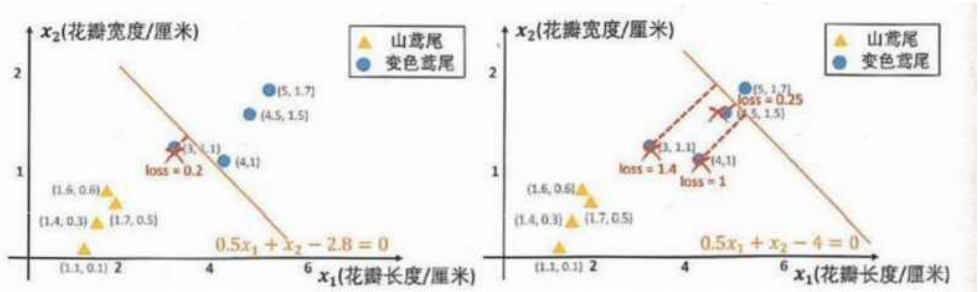


图 2-11 感知器在误分类数据上的损失函数值

有了损失函数衡量分类器对数据的误分类程度后，我们可以用优化的方法来调整分类器的参数，以减少分类器对数据的误分类。上面感知器学习算法中调整参数的规则是优化方法在感知器损失函数上的具体应用。细心的同学可能会发现，损失函数是在整个训练数据集上求得的，如果用它来更新参数，则是利用了整个数据集中被误分类的数据；而感知器学习算法中的第二步是每一次随机选取一个样本，如果是误分类

样本则用它来更新参数，这样不断迭代一直到训练数据中没有误分类数据为止。这是感知器损失函数利用优化方法得到感知器学习算法中做的一点小改动。

一般地，优化(optimization)就是调整分类器的参数，使得损失函数最小的过程。我们通过一个直观的例子来理解优化过程。为方便展示，我们只考虑两个参数：a_1, a_2。如图 2-12 所示，对于取值不同的每一组参数 a_1, a_2 ，都对应着一个损失函数的值。在左图中我们把所有这些损失函数值在三维坐标系中画出来，右图是对应的等高线图，即同一条等高线上具有相同的损失函数值。

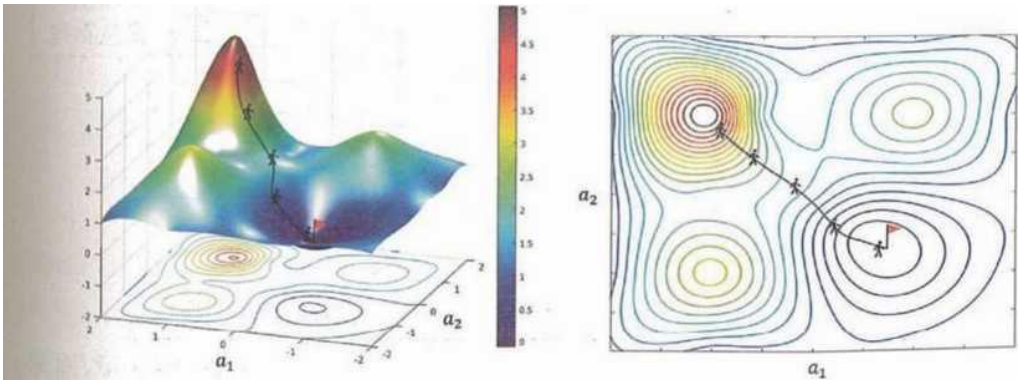


图 2-12　过程优化过程示意图

从图 2- 12 左图中可以看到损失函数值组成的曲面就像连绵起伏的山一样。山有高有低，有山峰有山谷。损失函数值最小的点就是海拔最低的山谷。优化的目标是使得损失函数的值最小，就是希望走到海拔最低的山谷。那么，优化的过程就是从山上走到山谷的下山过程。

如果我们每一次都沿着当前位置往下山的方向走一小步，这样就能保证每一小步后都能够走到海拔更低的位置，即得到更小的损失函数值，直到我们到达海拔最低的 山谷。此时我们便取得了最小的损失函数值。

---------- ----------

实验 2-1

1．利用上面介绍的感知器学习算法，训练一个感知器分类器。

2．修改感如器中初始化的参数，训练多个分类器，观察得到的分类直线是否一致。

3．　修改感知器学习算法中的学习率，训练多个分类器，观察得到的分类直线　是否一致。

支持向量机

在上面一节，我们介绍了感知器学习算法。通过实验，我们知道了在同样的训练数据下，感知器算法由于初始参数的选择不同或者学习率的不同等都会得到不同的分 类直线。这些不同的分类直线都能够将不同类别的数据分开，那么它们之间有没有优 劣之分呢？

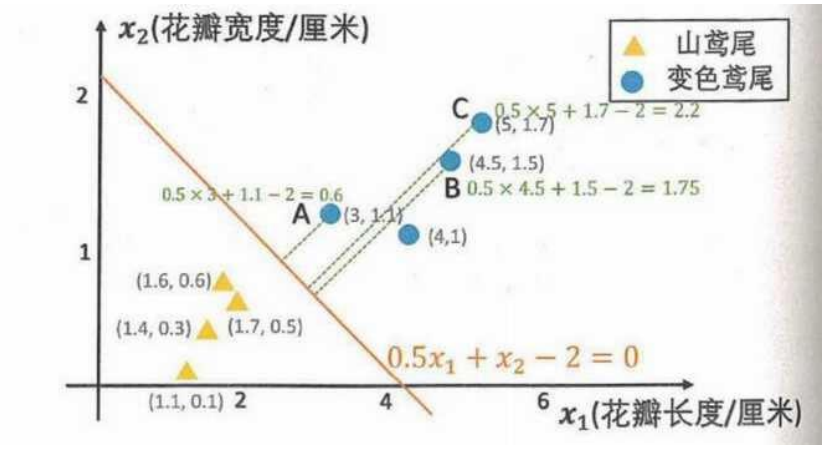


图 2-13 一个点距离分类直线 的远近可以表示分类预测的确 信程度

我们首先来看一个例子。在图 2-13 中，有 A、B, C 三个样本，它们都在分类直线的一侧。点 C 距离分类直线较远，如果预测该类为变色鸢尾，我们就比较确信这个预测是正确的；点 A 距离分类直线很近，如果分类器输出为变色鸢尾，我们就不那么确信这个预测了；点 B 在点 A 和点 C 之间，那么预测的确信度也在 A 和 C 之间。

一般地，一个点距离分类直线的远近可以表示我们对分类预测的确信程度如何表示点到分类直线的远近呢？在分类直线确定的情况下，$|a_1x_1+a_2x_2+b|$ 就能相对表示这个远近。而 $a_1x_1 + a_2x_2 + b$ 和 y 符号是否一致能够表示分类是否正确。所以， $y(a_1x_1+a_2x_2+b)$既表示了分类的正确性，又表示了预测的可信程度。图 2-13 中的 计算简单地佐证了这个结论。同学们会发现，这个式子也被用在感知器的损失函数中。

在上面的例子中我们固定了分类直线，分析不同数据点，直观地得到了“一个点距离分类直线越远，分类预测的可信程度越高”的结论。类似地，在给定一批训练数据后，我们希望训练得到的分类直线在分类准确的前提下，离开数据点越远越好，这样我们得到的分类器预测的可信程度相对较高。实际上，我们只要关注离分类直线最近的点的距离，使得它们距离分类直线越远越好。我们把两个类别中离分类直线最近的点到直线的距离和称为分类间隔(classification margin)。.

图 2-14 中画出了两条分类直线，它们都能正确分类山鸢尾和变色鸢尾。图中的阴影区域展示了对应分类器的分类间隔，可以看到，橙色直线的阴影区域更宽，对应 的分类间隔更大，其预测的确信程度也会越高。

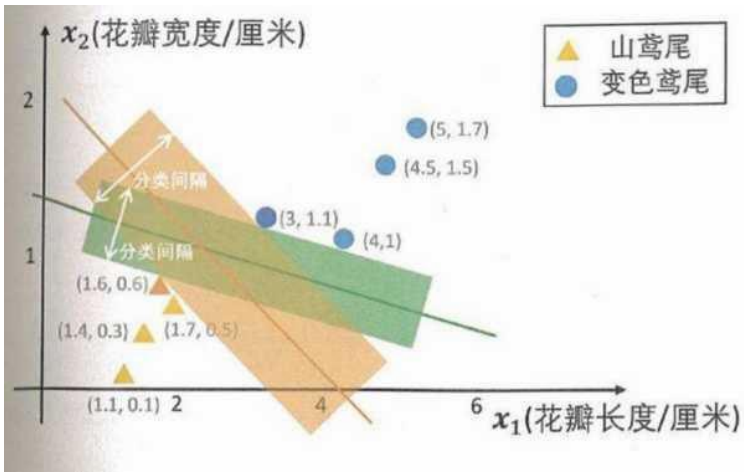


图 2-14　不同的分类直线有不同的分类间隔

支持向量机(support vector muiehine, SVM)是在特征空间上分类间隔最大的分类器，它与感知器一样. 是对两个类别进行分类。线性分类器是分类器中的一种，类似地，线性支持向量机也是支持向量机中的一种。若无特殊说明，我们这里说的支持向量机的是线性支持向量机。

直观上，我们很容易找到分类间隔最大的分类直线。同学们可以动手试试，想象自己拿了一根很粗的粉笔，要画一条线使得两类数据被正确分开，同时所画的线条是最粗的。那么所画的这条最粗的直线就有最大的分类间隔，对应着支持向量机要找的分类直线。

细心的同学会发现，在图 2-15 中画最粗的线，并不是和所有的数据点都有关系， 而只和一部分的点有关系。如果我们只保留图中和阴影区域相接触的点，去除其他所有数据点，我们可以使用同样的最粗线条来区分这两类数据，即得到的分类直线不变。事实上，我们把这样和阴影区域相接触的点称为支持向量(support vector)，这也是支持向量名字的由来。支持向量是那些能够定义分类直线的训练数据，也是那些最难 被分类的训练数据，直观地说，它们就是对求解分类任务最富有信息的数据。

实验 2-2

1. 利用提供的支持向量机学习算法，训练分类器。
2. 分别可视化由感知器和支持向量机训练出来的分类器的分类间隔，然后进

拓展阅读：支持向量机的损失函数

通过上面用粗粉笔画直线的例子，我们能够直观地找到最大分类间隔。支持向量机是分类间隔最大的分类器，那么我们能否像对感知器一样，写出支持向量机的损失函数，然后在训练数据上通过优化的方法对其进行求解呢?

答案是肯定的。我们接下来简单地介绍支持向最机损失函数的建立过程。我们在一个简单的情况下讨论：假定两类数据是同上面鸢尾花分类的例子一样，可以被一条直线分开。

我们的目的是确定分类直线 $a_1x_1 + a_2x_2 + b = 0$ 的参数 a_1，a_2，b。支持向量机的想法就是最大化分类间隔。如果数据点（x_1，x_2）被直线正确分类，那么这个点到直线的距离可以由下面的式子计算：

$$\frac{|a_1x_1 + a_2x_2 + b|}{\sqrt{a_1^2 + a_2^2}} = y \times \frac{a_1x_1 + a_2x_2 + b}{\sqrt{a_1^2 + a_2^2}}$$

我们由此可以定义训练数据中任一数据（$x_1^{(i)}$，$x_2^{(i)}$）和分类直线的几何间隔 $\gamma^{(i)}$ 为：

$$\gamma^{(i)} = y^{(i)} \times \frac{a_1x_1^{(i)} + a_2x_2^{(i)} + b}{\sqrt{a_1^2 + a_2^2}}$$

几何间隔和我们说的点到直线的距离有什么关系呢？如果数据点被正确分类，几何间隔就是点到直线的距离；如果没有正确分类，它们相差一个正负符号，所以数据点到分类直线的几何间隔一般是点到直线的带符号的距离。

我们可以定义全部训练数据到直线的几何间隔为所有数据点到直线的几何间隔的最小值，即

$$\gamma = \min_{i=1,\cdots,N} \gamma^{(i)}$$

min 符号表示最小化后面表达式的值。从图 2－15 中，我们可以看到分类间隔是几何间隔的两倍，为 2γ。

因此，最大化分类间隔就是最大化 2γ，用数学符号表示为 $\max_{a_1,a_2,b} 2\gamma$，max 符号表示最大化后面表达式的值，max 下面的 a_1，a_2，b 表示可以改变的参数。最大化 2γ 等价于最小化 $\frac{2}{\gamma}$，用数学符号表示为 $\min_{a_1,a_2,b} \frac{2}{\gamma}$。这个形式可以看成是支持向量机的损失函数，我们希望最小化这个损失函数。

这个问题的求解还需要保证每个训练数据点到分类直线的几何间隔至少是 γ。这样，整个优化问题就可以表示为：

$$\min_{a_1,a_2,b} \frac{2}{\gamma}$$

同时满足对每一个 i，$y^{(i)} \times \frac{a_1x_1^{(i)} + a_2x_2^{(i)} + b}{\sqrt{a_1^2 + a_2^2}} \geqslant \gamma$

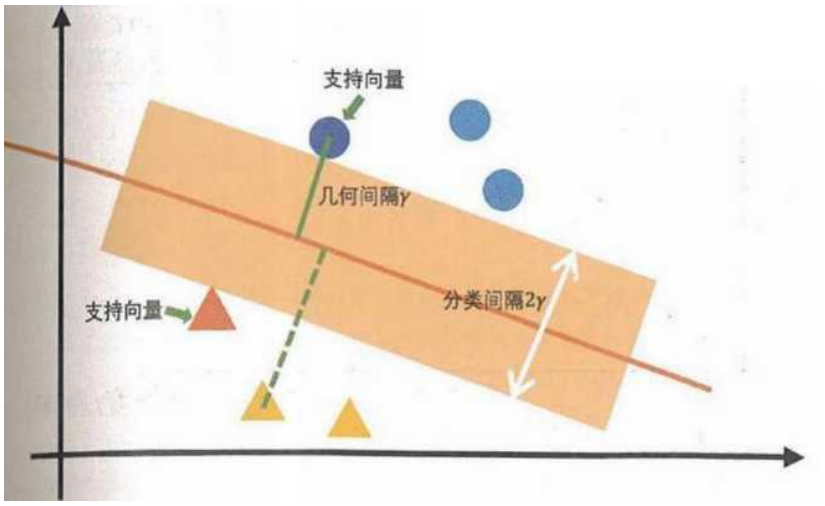


图 2-15 几何间隔和分类间隔

这样我们就得到了支持向量机完整的损失函数。

这个问题可以用优化的方法求解，但具体过程已经超出了同学们现阶段的知识储备，等到同学们以后学习了相关的数学工具后，就可以求解了。

2.4 实践出真知：测试和应用

我们已经了解了两种分类器的学习算法：感知器与支持向量机。在得到分类器之后，我们希望知道分类器的分类效果怎么样，哪一个学习算法获得的分类器性能最好。于是，我们需要测试的环节。

测试就像我们学习之后的考试。在考试中，同学们一般会面对一张试卷进行答题；答题结束后，老师进行阅卷，最终给出评分。

类似地，在分类器的测试阶段，它会面对一批测试数据并要对每一个测试样本做出预测结果。如果分类的结果和测试样本的标注一样，那么分类正确，否则分类错误。比如在区分鸢尾花品种的例子中，测试数据中有一朵鸢尾花，它的花瓣长度是 1.5 厘米，宽度是 0.4 厘米。将测试样本的特征向量（1.5, 0.4) 画在特征空间中，得到图 2-16 中红色五角星位置。可以看到它位于分类直线山鸢尾的一侧，预测为山鸢尾。而这朵鸢尾花确实为山鸢尾，所以分类正确。

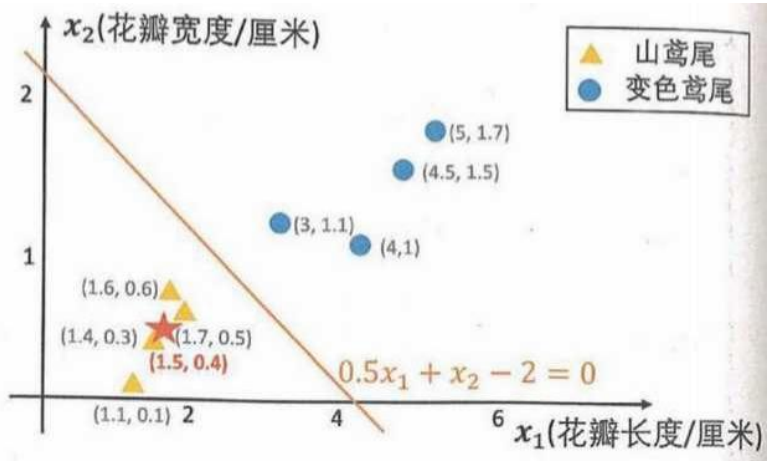


图 2-16 测试

在整个测试集上都测试一遍后，我们统计出分类器分类正确的样本数，它与测试样才总数的比率可以反映预测的准确程度，被称为分类准确率(classification accura- cy)--它相当于老师进行阅卷后给出的分数。

$$\text{分类准确率} = \frac{\text{分类正确的样本数}}{\text{测试样本总数}} \times 100\%$$

知道了分类准确率后，我们便知道了分类器的效果，这样就可以在多个分类器之间进行比较，选择一个最好的分类器。

经过测试选择了一个最佳的分类器，接下来就是它大显身手的时候了。如果我们看到了一朵鸢尾花，想知道它属于哪个品种，只需取出尺子量一下它的花瓣长度和宽度，然后输入到训练好的分类器中，分类器就会输出它的预测结果。这个过程就是分类器的应用阶段。

图 2-17 训练数据、测试数据和应用阶段的数据

测试和应用有什么区别呢？从作用上讲，测试是用来评判分类器表现好坏，选择最优分卷器；应用则是分类器在实际情况中的使用。我们对一个问题一般会训练多个分类器，在测试数据上进行测试，选择表现最好的一个分类器，通过测试后再拿到实际中应用，从数据角度讲，测试阶段使用的是预先采集并且标注好的测试集；而在应用阶段，数据是从实际中来的、没有标注过的，并且更为复杂多变，如图 2-17 所示。

实验 2-3

1. 测试：由感知器和支持向量机训练算法得到的分类器的分类准确率。
2. 应用：输入鸢尾花的花瓣长度和宽度. 得到分类器的预测结果。

2.5 五花八门：多类别分类

上面我们介绍了对两类物体的分类，即二分类(binary classification)。实际中， 我们往往需要对多种类别进行分类，比如区分牡丹、荷花、梅花等多种花。那么我们解决这样的多分类(mulliclass classification)问题呢？

回顾前面的例子. 我们使用了一个分类函数解决了二分类问题。在多分类问题中，我们是否可以把多分类问题转化为多个二分类问题呢？我们使用多个二分类函数, 其中的每一个分类函数都有自己的参数，只负责区分一个类别。如图 2-18 所示，我们有三个分类器，分别是牡丹、荷花、梅花的分类器，它们只负责区分某一个类别是牡丹不是牡丹，是荷花不是荷花，是梅花不是梅花。当输入一张图片的特征向最后，三个分类器都能够输出自己的预测，综合三个预测结果，我们最终就能够得到多分类的预测结果。

具体地说，如果 f_1输出为正，f_2和 f_3输出为负，那么我们可以确定地说类别是牡丹。但是不排除有两个分类器输出都可能为正的情况，这个时候该怎么综合判断呢？这实际上是一个不确定的问题，就像我们平常生活中会听到天气预报说明天的降雨概 率为 80%——明天不一定会降雨，但是会有很大可能性下雨。那么我们是否也可采 用类似的策略呢？

事实上，我们可以将分类器函数 f_1,f_2和 f_3的输出值通过一个归一化指数函数。 它可以把输出转变为概率——说明输入物体属于某一类的可能性。归一化指数函数在下一章的神经网络多分类问题中被广泛应用。

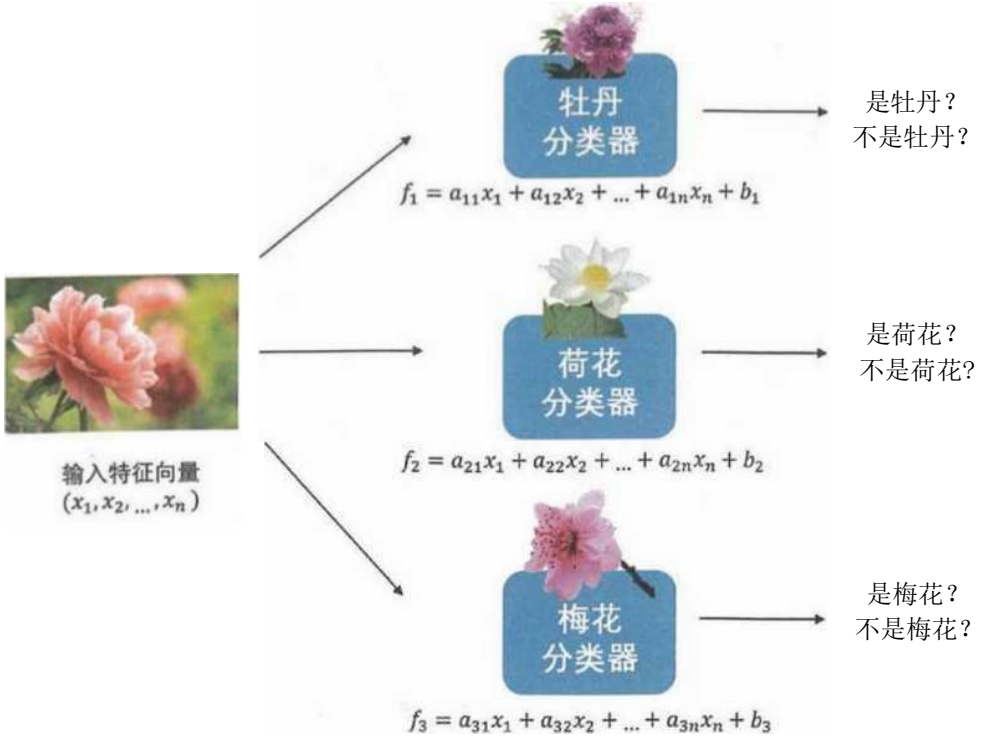


图 2-18 多分类器示意图

归一化指数函函数(softmax)是将一个向量（比如多个二分类函数的输出值可以组成一个向量）“压缩”到另一个向量中，使得其中每一个元素的范范围(0，1)之 间，并且所有的元素和为 1。这个过程就被称为归一化。softmax 函数具体的形式由下式给出：

$$\sigma(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

其中，j=l，…，*K*。这个向量总共有 K 项，首先对它们分别进行指数变换，然后计算每个指数变换后的输出占所有输出总和的比例。因为要先进行指数变换，然后 归一化，所以 softmax 函数被称为归一化指数函数。

表 2-1 列举了归一化指数函数的计算。如果三个二分类器分别输出-1, 2, 3, 组成向量（-1, 2, 3)。经过归一化指数函数后，得到向量（0.013, 0.265, 0.722)，它们加起来为 1。输出向量中最大值 0.722 对应着输入向量中的最大值 3 这个函数通常的意义是对向量进行归一化，凸显其中最大的值并抑制远低于最大值喝其他值。

表 2-1　归一化指数函数计算

输入：	-1	2	3	求和
指数变换e^{z}	$e^{-1}\approx 0.368$	$e^{2}\approx 7.389$	$e^{3}\approx 20.086$	27.843
归一化指数函数	$\frac{0.368}{27.843}\approx 0.013$	$\frac{7.389}{27.843}\approx 0.265$	$\frac{20.086}{27.843}\approx 0.722$	1

经过归一化指数函数后，它们的值都大于零，并且加起来和为 1。我们可以将它们看作概率，即输入属于某一类的可能性大小。比如输出(0.013 , 0.265 , 0.722) 可以表示有 1.3%的可能性属于第一类，有 26.5%的可能性属于第二类，有 72.2%的 可能性属于第三类。因此输入很有可能属于第三类。

通过归一化指数函数后，分类器输出的值有了更深一层的含义，不仅能够告诉我们输出的类别，同时能够将我们之前直观的确信程度转变成可以衡量的概率，即分类器对这个结果的预测把握。比如输出概率为 99%的牡丹分类，那么这个预测就很准确；如果输出概率为 65%的梅花分类，那么这个预测就没有那么大把握了，我们是 否分类器的预测，就需要再斟酌一下。

2.6 大显身手：二分类在生活中的应用

这一章我们介绍了二分类的问题--把事物分类成两个类别。二分类的问题在实 际生活中有着广泛的应用。生活中遇到的“是不是问题”都属于二分类的范畴一这是不是一张人脸？这是不是有癌症的医学影像？这是不是一处可能有矿藏的地方？.....下面我们介绍相机中的人脸检测和医疗中的癌症检测，来具体看看二分类技术是怎么应用在实际生活中的。

相机中的人脸检测

同学们在外出游玩的时候，都会拍照留念。无论是用手机还是单反，当镜头对准人脸的时候，都会出现一个矩形框，框出人脸的区域，如图 2-19 所示。那么，这个技术是如何做到的呢？

图 2-19 相机中的人脸检测

相机中的人脸检测技术使用的就是二分类技术。它的整个流程如图 2-20 所示， 一张照片首先被切割成一块块的图像块。这样的切割很密集，它们重叠连续地将照片切割成小的图像块。一张照片往往会有成千上万的图像块被切割出来。

然后每一个图像块都会经过人脸分类器去判别是否是人脸。人脸分类器是预先训练好的二分类器，就像我们训练区分鸢尾花品种的分类器一样。对于预测是人脸的图 像块，机就在这个图像块上出框的位置——这就是相机中的人脸检测的奥秘。

同学们可能会有疑问，拍照时，由于距离远近等导致人脸大小不一样怎么办？ 实际上，在切割图像块的时候，并不是只有一种尺寸的图像块切割，而是从小到大 有很多尺寸，这保证了能够涵盖几乎各种大小的人脸。所以，一张照片截取的图像 块数量是非常大的，由干使用的人脸分类器简单、运算量小，还有其他优化速度的 技术的使用，整个流程的时间往往很难被我们感知到，好像相机中的人脸检测是实 时的一样。

图 2-20 相机中的人脸检测流程

同学们可能还有疑问，不同尺寸、不同位置的图像块可能同时都被判别为是人脸，那么就会有很多重叠框了。事实上确实如此，像图 2-21 左图一样，在人脸附近

截取出来的不同位置、不同尺寸的图像块都是人脸。这些框都在人脸的附近，我们可 以通过后处理融合技术，将这些框融合为一个框，得到右图的结果。

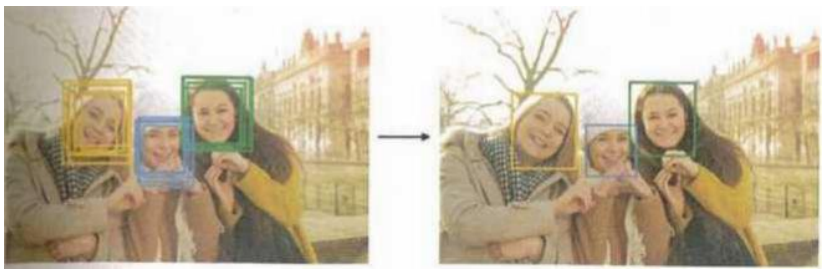


图 2-21 脸检测中多个框的 融合

癌症检测

根据患者的生物组织样本图像.判断是否患有癌症是二分类在医疗领域的一个应用。病理学家在检查患者的生物组织样本后，能够诊断是否患有癌症。他们的诊断对于患者的治疗至关重要。然而，病理切片的审查非常复杂，需要多年的培训和丰富的专业知识及经验。

癌症检测，从分类的角度来看.就是一个二分类问题一判断生物组织样本的每一个区城是否是肿瘤(tumor)。目前，随着人工智能系统的进步，在一些癌症的诊断 上，人工智能系统诊断癌症的准确率正在逐步向有经验的病理学家靠拢。图 2-22 中，左图是淋巴结(lymph node)活检(biopsy)的图像，右图是谷歌公司最近的检测结果，这个结果在测试数据上已经能够达到并超过一般的人类病理学家。这个结果 是振奋人心的，但实际中癌症的组织样本更加复杂，病理图像会受到其他因素的影 响，甚至有可能出现罕见的病理切片图像，它们不曾在训练数据中出现过，人类病理 学家可以依靠经验和知识处理这些状况。而人工智能诊断家们在这方面还有较大

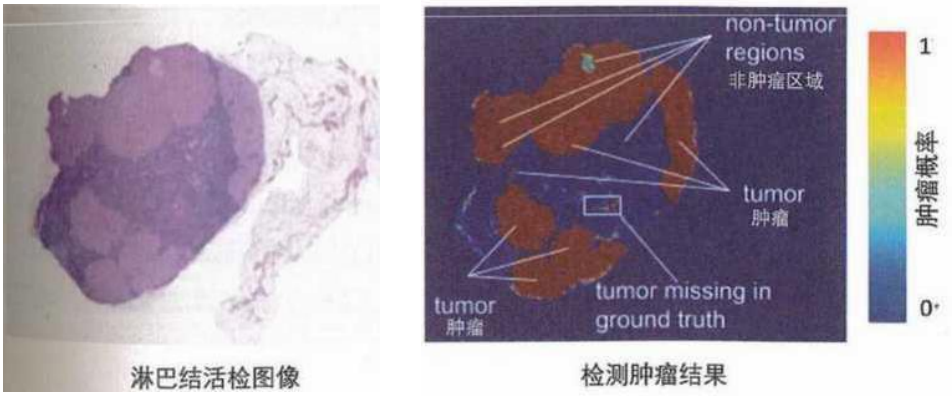


图 2-22 由炳理切片图像检测癌底

在癌症检测中，良性和恶性往往很难区分，比如在图 2-23 中是一块包含乳腺癌 转移的恶性肿瘤和一块正常的组织——巨噬细胞，它们外观上十分相似，目前的人工智能系统能够比较正确区分肿瘤和巨噬细胞(macrophage) ——它们向着有经验的人 类病理学家的水平又往前迈进了一步。

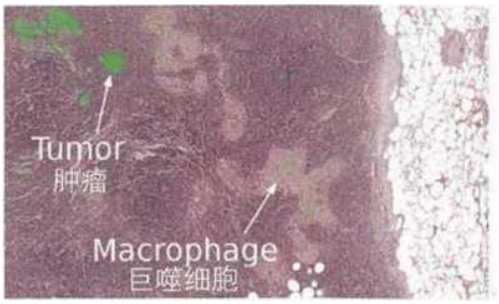


图 2-23 正确区分淋巴结中的 肿瘤和巨噬细胞

这些目前还在实验室的人工智能诊断家们，正在逐步走出去，辅助医生的诊断， 帮助做出更准确、更及时的癌症诊断。这不仅提高了效率，减轻了医生负担；同时也使得检测结果更加准确，帮患者争取了宝贵的治疗时机。技术在不断进步，可以预想到不远的未来，人工智能技术将大量地运用在医疗领域。

2.7 本章小结

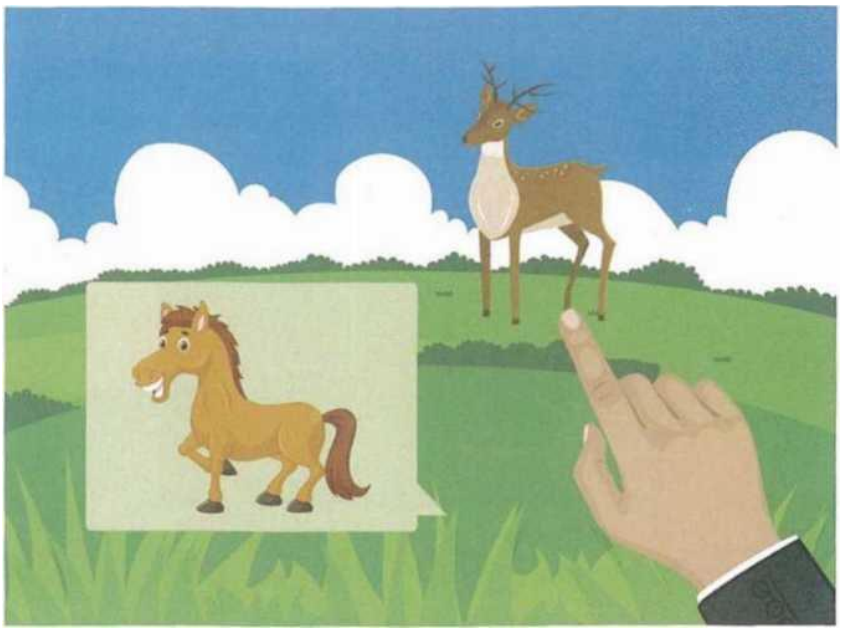
分类是把事物归属到它所属类别的过程，在生活中有着广泛应用。特征和分类器是分类中的重要概念。特征是根据事物自身的特点，提取的某方面数字或属性，它可以用特征向量来表示。而分类器是从特征向量到类别的函数。

分类过程可以分为三个阶段---征提取、分类器的训练以及测试应用。特征提取是由数据到特征向量的过程，是传统分类方法中的重点。得到特征向量之后，我们使用数据和算法来训练分类器。分类器通过测试后，就可以应用在实际生活中了。

分类器的训练是由训练算法来完成的，不同的训练算法可能会得到不同的分类器。 本章介绍了感知器和支持向量机训练算法。它们都有着自己的损失函数。损失函数可以衡量分类器在训练过程中输出错误的程度，然后通过优化方法就能够得到分类器了。

第三章　别具慧眼：识图认物

铭铭喜欢摄影，相册里也已经积累了无数的照片，有的是动物园中可爱的生灵，有的是山野里缤纷的花，有小猫、小狗，也有汽车、飞机。 在翻看这些照片时，很多细节已不能记起。如这 只可爱的动物名字是什么？这株灵动的花儿叫什么？小伙伴抱着的宠物狗是什么品种？“要是电脑能自动识别这些物体，能告诉我就好了。”*铭*铭心里想。



《史记》中记载有奸臣赵高指鹿为马，混淆是非；

《艾子杂说》中，有人欲以鹘（hu,鹰属猛禽）猎兔而不识鹘, 买凫(fu, 鸭 子）而去，逼凫捉兔，成为笑谈。

此时，若有一宝，可以不受人为因素的影响，准确辨别动物种类，是不是这些难题就可以化解了呢？

千年之后的今天，深度学习技术的出现正好为我们提供了这样一个契机，让这样的奇思妙想成为可能。下面，让我们一起揭开深度学习技术的神秘面纱，开启亲手制 作这件“明察秋毫之宝”的神奇旅程。

3.1 温故知新：基于手工特征的图像分类

铭铭的相册中有小猫、小狗，也有汽车、飞机。但是，里面也有一些一眼不是很确定的事物的照片。比如图 3-1 中的第一张是企鹅还是别的什么鸟呢？第二张照片是什么种类的猫呢？识别照片中的物体是什么类别是一个分类任务。通过第二章的学习，我们知道分类任务包含两个核心步骤：特征提取与特征分类。如图 3-2 所示，在鸢尾花分类的例子中，我们通过测量花瓣的长和宽，从一个鸢尾花样本中提取一个二维的特征向量。随后，这个特征向量被输入到分类器，经过一系列计算，分类器就

可以判断出这朵鸢尾花的类别。我们可以遵循同样的流程，设计一个用于对图片进行分类的系统。那么对于图片分类这个任务，我们应该使用什么样的特征？怎样从图片 中有效提取它们呢？在回答这些问题之前，我们先了解一下计算机眼中的图片是什么 样的。

图 3-1 铭铭的相册

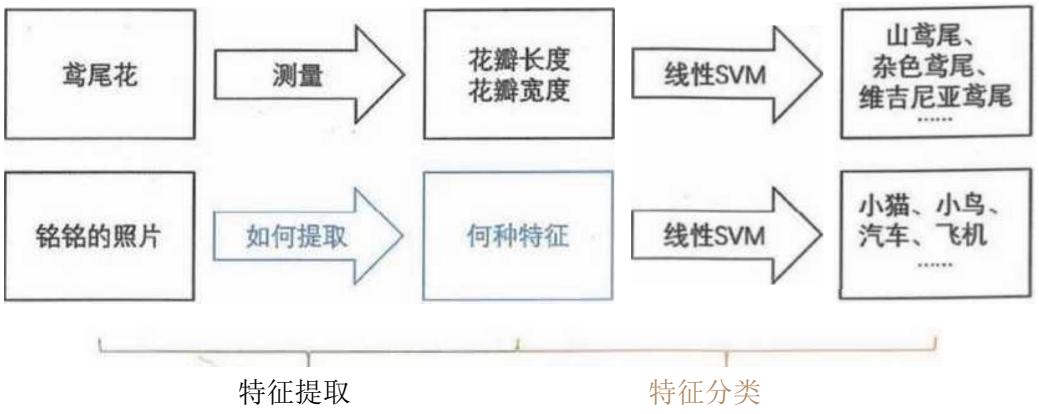


图 3-2 分类任务的两个基本步骤

计算机眼中的图像

在学习图像特征提取之前，我们先来看一下图像（image)在计算机中是如何表示的。如图 3-3,如果将一幅图像放大，我们可以看到它是由一个个的小格子组成的，每个小格子是一个色块。如果我们用不同的数字来表示不同的颜色，图像就可以表示为一个由数字组成的矩形阵列，称为矩阵（matrix）,这样就可以在计算机中存储。这里的小格子我们称之为像素（pixel）;而格子的行数与列数，统称为分辨率 (resolution)。我们常说的某幅图像的分辨率是 1280 x 720，指的就是这张图是由 1280 行、720 列的像素组成的。反过来，如果给出一个数字组成的矩阵，我们将矩阵中的每个数值转换为对应的颜色，并在电脑屏幕上显示出来，就可以复现这张图像。

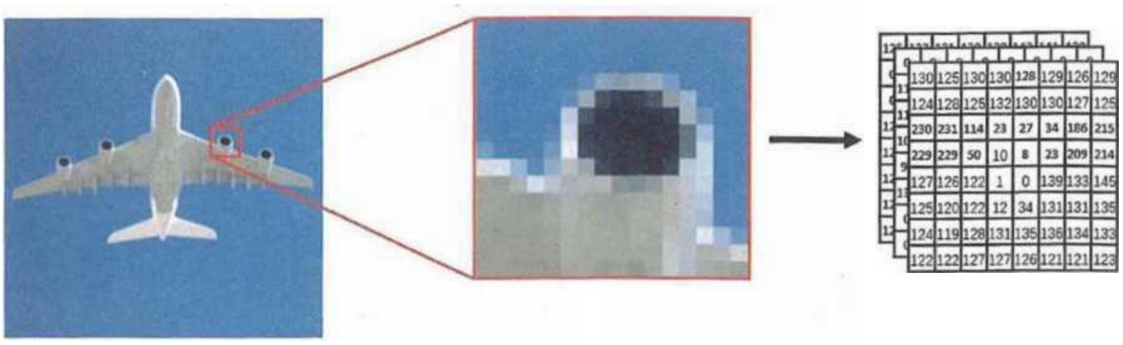


图 3-3　计算机中图像的表示

照片分黑白和彩色。在图像里我们相应地有灰度图像和彩色图像。对于灰度图像，由于只有明暗的区别，因此只需要一个数字就可以表示出不同的灰度。通常我们用 0 表示最暗的黑色，255 表示最亮的白色，介于 0 和 255 之间的整数则表示不同明 暗程度的灰色。对于彩色图像，我们用(R. G, B)三个数字来表示一个颜色，它 表示用红(R)、绿(G)、蓝(B)三种基本颜色亜加后的颜色。对于每种基本颜色，我们也用介于 0 到 255 之间的整数表示这个颜色分量的明暗程度。如图 3-4 所示， 三个数字中对应某种基本颜色的数字越大，表示该基本颜色的比例越大. 例如， (255, 0, 0) 表示纯红色，(0，255，0) 表示纯绿色，(135，206，255)是天蓝色。

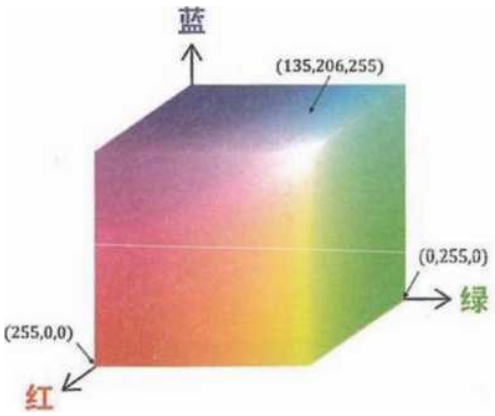


图 3-4 色彩的表示

我们现在知道，一张彩色图像可以用一个由整数组成的立方体阵列来表示。我们称这样按立方体排列的数字阵列为三阶张量(tensor) 。这个三阶张量的长度与宽度即 为图像的分辨率，高度为 3。对数字图像而言，三阶张量的高度也称为通道(chan- nel) 数，因此我们也说彩色图像有三个通道。矩阵可以看作是高度为 1 的三阶张量，因此灰度图像只有一个通道。

知识链接：张量

张盘是数学、物理及工程等学科中的-个基本概念，我们之前遇到的许多概念都是张量的特殊形式，例如标量（scalar）属于零阶张量，向量是一阶张量，而矩阵则是二阶张量。

图像特征概述

在正式学习图像特征之前，我们可以先简单思考一下，什么样的特征可以区分这 些照片呢？例如在表 3-1 中，我们将“有没有翅膀”作为一个特征，就可以区分小 鸟和小猫，也可以区分汽车和飞机。再将“有没有眼睛”作为另一个特征，我们就可以完美地区分这四类照片了。

表 3-1 可以区分四类照片的特征

	小猫	小鸟	飞机	汽车
特征 1：有没有翅膀	没有	有	有	没有
特征 2：有没有眼睛	有	有	没有	没有

那么怎样从图像中提取这两个特征呢？对干人类而言，这个过程非常简单，我们只要看一眼图片，大脑就可以获取这些特征。但是对于计算机而言，一幅图片就是以 特定方式存储的一串数据。让计算机通过一系列计算，从这些数据中提取类似“有没有翅膀”这样的特征是一件极其困难的事情（如图 3-5 所示）。

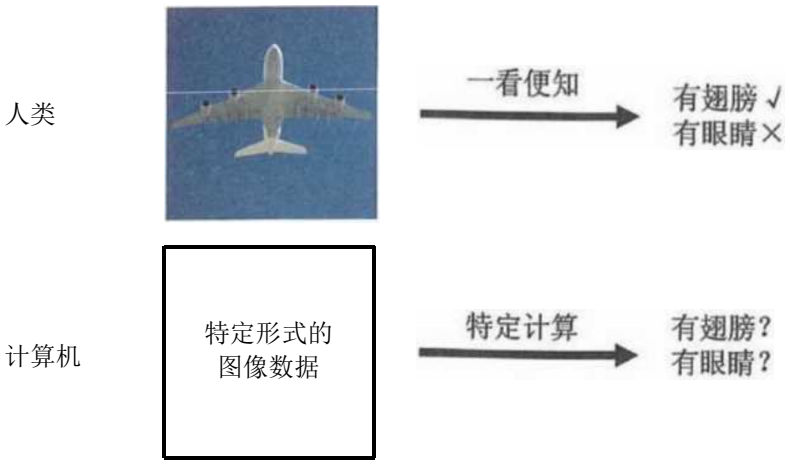


图 3-5 人类与计算机在获取图像特征上的区别

在深度学习(deep learning)出现之前，图像特征的设计一直是计算机视觉 (computer vision)领域中一个重要的研究课题。在这个领域发展的初期，人们手工设计了各种图像特征，这些特征可以描述图像的颜色、边缘(edge)、纹理(lexture) 等基本性质，结合机器学习技术，能解决物体识别(object recognition)和物体检测 (object detection)等实际问题。

既然图像在计算机中可以表示成三阶张量，那么从图像中提取特征便是对这个三阶张量进行运算的过程。其中非常重要的一种运算是卷积。

卷积运算

卷积运算在图像处理以及其他许多领域有着广泛的应用。卷积和加减乘除一样， 是一种数学运算。参与卷积运算的可以是向量、矩阵或三阶张量。我们先从向量的卷积入手，讲解卷积的基本步骤，再将其推广到矩阵和三阶张量。

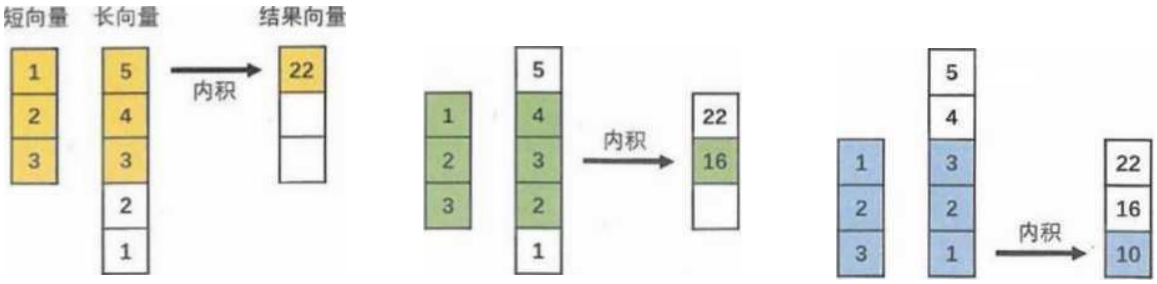


图 3-6 向量的卷积

两个向量卷积(convolution)的结果仍然是一个向量。它的计算过程如图 3-6 所示。我们首先将两个向址的第一个元素对齐，并截去长向量中多余的元素，然后，我们计算这两个维数相同的向量的内积，并将算得的结果作为结果向量的第一个元素。接下来，我们将短向量向下滑动一个元素，从原始的长向量中截去不能与之对应的元 素，并计算内积。重复“滑动-截取-计算内积”这个过程，直到短向量的最后一个元素与长向量的最后一个元素对齐为止。最后就可以得到这两个向量卷积的结果。作为一种特殊情形，当两个向量的长度相同时，不需要进行滑动操作，卷积结果是长度为 1 的向量，结果向量中这个元素就是两个向量的内积。

知识链接：向量间卷积的数学描述

向量之间卷积运算的过程可以数学的语言来描述。对于维数为 m 的向量 $a = (a_1, a_2, \ldots, a_m)$ 和维数为 $n(n \geqslant m)$ 的向量 $b = (b_1, b_2, \ldots, b_n)$，二者卷积运算的结果是一个维数为 $n - m + 1$ 的向量 $c = (c_1, c_2, \ldots, c_{n-m+1})$，并满足对于任意 $i \in \{1, 2, \ldots, n - m + 1\}$，有 $c_i = \sum_{i=1}^{m} a_i b_{i+i-1} = a_1 b_i + a_2 b_{i+1} + \ldots + a_m b_{i+m-1}$。通常我们使用符号“ * ”来表示卷积运算，例如上述例子中的卷积运算可以表示为：$(1,2,3) * (5,4,3,2,1) = (22,16,10)$。

从上面的定义可知，卷积结果的维数通常比长向量低。有时候我们为了使得卷积之后维数和长向量一致，会在长向量的两端补上一些 0。对于图 3-6 中的例子，我们 可以把长向量的两端各补一个 0, 变成(0, 5, 4, 3, 2, 1, 0), 再进行卷积运算，就可以得到维数仍然为 5 的结果向量。

类似地，我们可以定义矩阵的卷积。在此之前. 我们首先需要将内积运算拓展到 矩阵上。如图 3-7 所示，对于两个形状相同的矩阵，它们的内积是每个对应位置的 数字相乘之后的和。

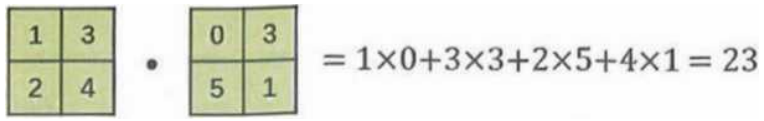


图 3-7 矩阵的内积

进行向量的卷积时(图 3-6), 我们只需要沿着一个方向进行滑动；而进行矩阵的卷积时(图 3-8), 我们需要沿着横向和纵向两个方向进行滑动。

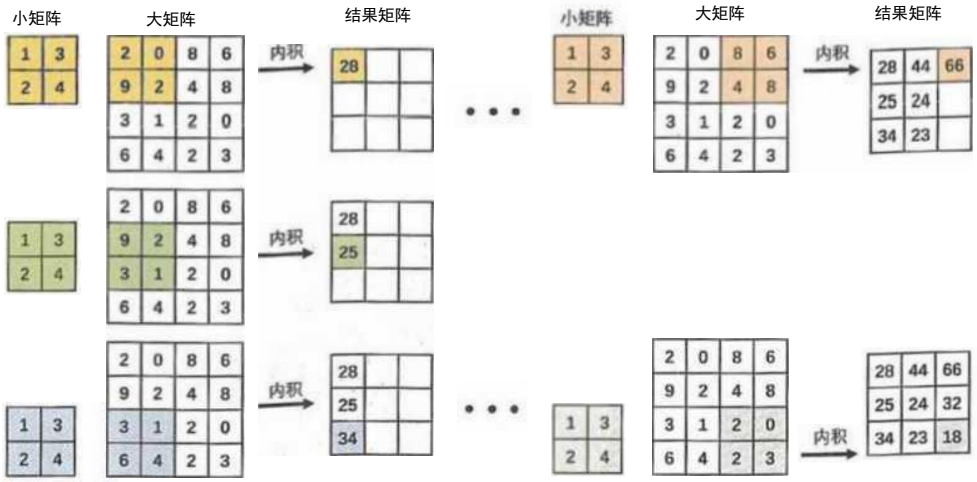


图 3-8 矩阵的卷积

类似地可以定义三阶张量之间的卷积（如图 3-9 所示）。这里我们只讨论一种简单的情形，有兴趣的同学可以 探究一下三阶张量卷积的一般形式。当两个张量的通道数相同时，滑动操作和矩阵卷积一样，只需要在长和宽两个方向进行。卷积的结果是一个通道数为 1 的三阶张量。

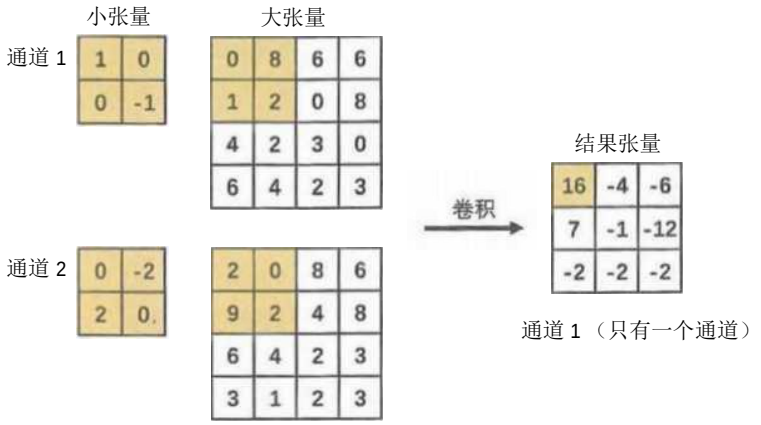


图 3-9 三阶张量的卷积

利用卷积提取图像特征

卷积运算在图像处理中应用十分广泛，许多图像特征提取方法都会用到卷积。以灰度图为例，我们知道在计算机中一幅灰度图像被表示为一个整数的矩阵。如果我们用一个形状较小的矩阵和这个图像矩阵做卷积运算，就可以 得到一个新的矩阵，这个新的矩阵可以看作是一幅新的图像。换句话说，通过卷积运算，我们可以将原图像变换为 一幅新图像一这幅新图像有时候比原图像更清楚地表示了某些性质，我们就可以把它当作原图像的一个特征。这里 用到的小矩阵就称为卷积核（convolution kemel)。通常，图像矩阵中的元素都是介于 0 到 255 的整数但卷积，但卷积核中的 元素可以是任意实数。

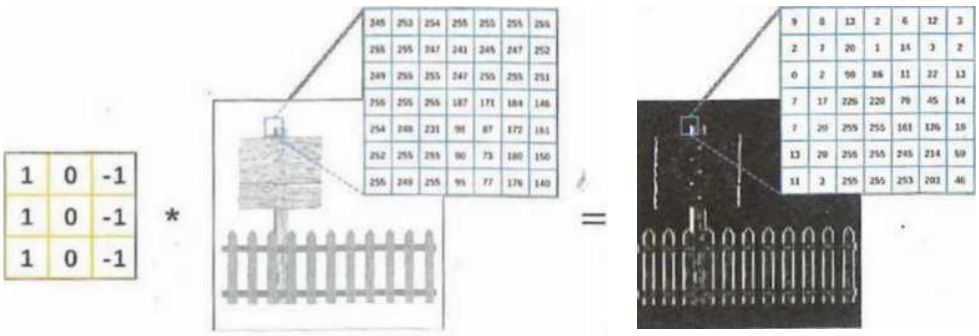


图 3-10 使用卷积提取竖向边缘

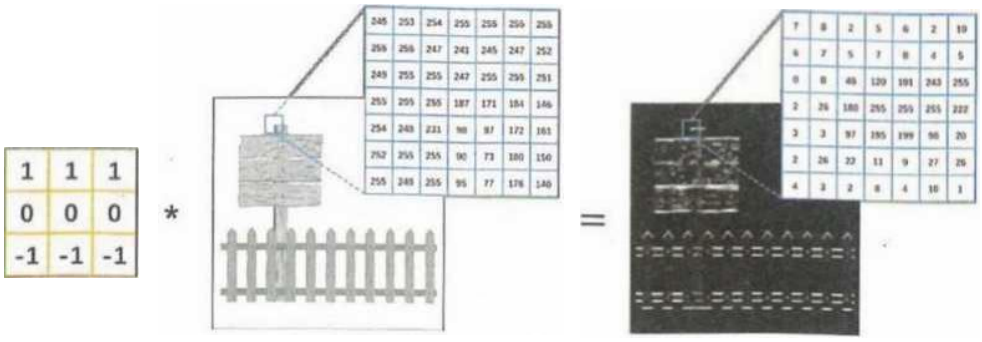


图 3-11 使用卷积提取横向边缘

通过卷积，我们可以从图像中提取边缘特征。在没有边缘的比较平坦的区域，图像像素值的变化比较小，而横向边缘上下两侧的像素差异明显，竖向边缘左右两侧的像素也会有较大差别。在图 3-10 的例子中，我们用三列 1、 0、-1 组成的卷积核与原图像进行卷积运算，可以从图像中提取出竖向边缘，在图 3-11 的例子中，我们用三行 1、 0、-1 组成的卷积核，从图像中提取出了横向边缘。事实上，这两个卷积核分别计算了原图像上每个 3x3 区域内左 右像素或上下像素的差值（为了将运算结果以图像的形式展示出来，我们对运算结果取了绝对值）。通过这样的减法运算，我们就可以从图像中提取出不同的边缘特征。

更进一步地，研究者们设计了一些更加复杂而有效的特征。方向梯度直方图（histogram of oriented graclienls, HOG）是一种经典的图像特征，在物体识别和物体检测中有较好的应用。方向梯度直方图使用边缘检测技术和一些 统计学方法，可以表示出图像中物体的轮廓。由于不同的物体轮廓有所不同，因此我们可以利用方向梯度直方图特 征区分图像中不同的物体（如图 3-12 所示）。

方向梯度直方图的提取过程主要包括两个步骤。首先我们利用卷积运算从图像中提取出边缘特征. 接下来，我们将图片划分成若干区域，并对边缘特征按照方向和幅度进行统计，并形成直方图一最后我们将所有区域内的直方图拼接起来，就形成了特征向量。具体过程相对复杂，我们在这里略去。

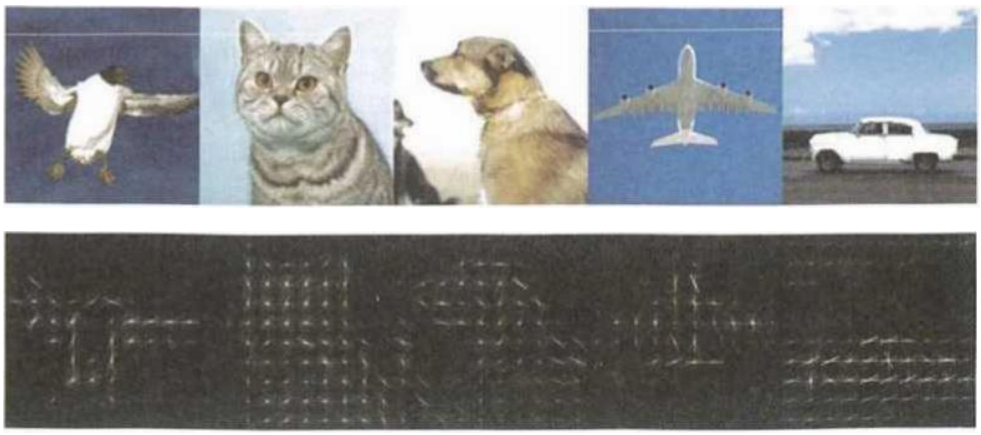


图 3-12 不同形状物体的方向梯度直方图

实验 3-1：利用图像待征进行图片分类
在这个实验中，我们将利用方向梯度直方图和多类支持向量机分类器，在 CIFAR 10 数据集上完成图片分类的任务。CIFAR 10 是一个常用的图像分类数据 集，由加拿大高等研究院的人工智能科学家搜集编排而成。该散据集包含 10 类 物体的图片，总计 6 万张。 1. 观察 CIFAR 10 数据集，对不同类别的图片有一个基本的认识，并区分开训练集与测试集。 2. 利用工具包提供的函数对全部图片提取方向梯度直方图。利用我们提供的工具包数对方向梯度直方图进行可视化，理解方向梯度直方图是对物体轮廓的 描述。 3. 利用生训练集上提取的方向梯度直方图完成支持向量机分类器的训练，记录训练集上的分类正确率。 4. 利用训练好的支持向量机分类器对测试集的方向梯度直方图进行分类，记录测试集上的分类正确率。

3.2 另辟蹊径：深度神经网络的图像分类

从特征设计到特征学习

通过上一节的学习和实验，我们学会了利用方向梯度直方图特征和支持向量机分类器完成图像分类的任务，然而分类的正确率并不太令人满意。事实上，这也是当时计算机视觉领域面临的一个问题：利用人工设计的图像特征，图像分类的准确率已经 达到“瓶颈”。

Image Net 挑战赛是计算机视觉领域的世界级竞赛，比赛的任务之一就是让计算机自动完成对 1000 类图片的分类。在 2010 年首届 Image Net 挑战赛上，冠军团队使用两种手工设计的特征，配合支持向量机，取得了 28.2%的分类错误率。在 2011 年 的比赛中，得益于更好的特征设计，第一名的分类链误率降低到了 25.7%。然而对于人类而言，这样的“人工智能系统”还远远称不上“智能”。如果我们将竞赛用的数据集交给人类进行学习和识别，人类的分类错误率只有 5-1%，低出当时最先进的

分类系统足足有 20 个百分点（如图 3－13 所示）。

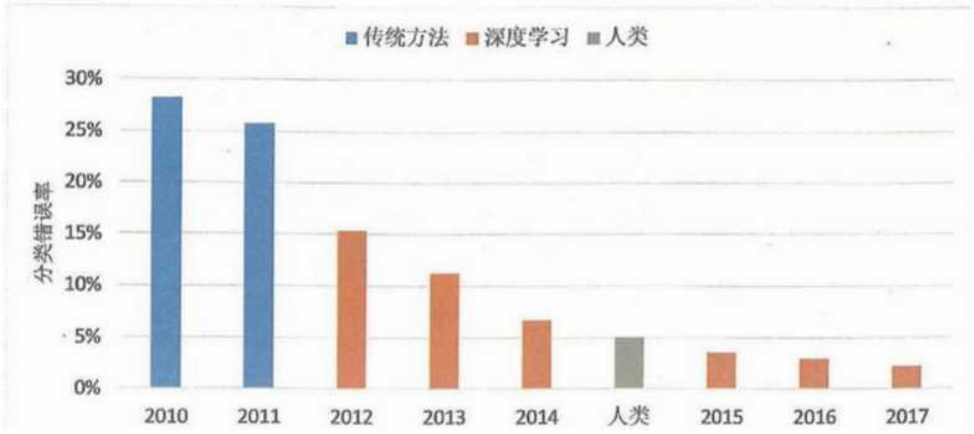


图 3-13 Image Net 挑战赛历年成绩

我们能否尝试提出更好的图像特征呢？或许可以。但这项工作往往需要领域内的兼具专业知识和创造力的科学家与工程师经过数年的摸索与尝试，甚至还需要一些运气成分才可能有所突破。特征设计的困难也极大地拖慢了计算机视觉的发展。

然而 2012 年的 Image Net 挑战赛给人们带来了惊喜，来自多伦多大学的参赛团队首次使用深度神经网络，将图片分类的错误率一举降低了 10 个百分点，正确率达到 84.7%。自此以后，Image Net 挑战赛就是深度神经网络比拼的舞台。仅三年后，来自微软研究院的团队提出一种新的网络结构，将错误率降低到了 4.9%, 首次超过了人类的正确率。到了 2017 年，图片分类的错误率已经可以达到 2.3%。这是举办 Image Net 挑战赛的最后一年，因为深度神经网络已经比较好地解决图片分类的问题。

深度神经网络之所以有这么强大的能力，就是因为它可以自动从图像中学习有效 的特征。在图像分类的任务中，手工设计的特征往往很难直接表达“有没有翅膀 ” 或"有没有眼睛”这样高层次的抽象概念。然而深度神经网络出现之后，这一切便成为可能。在计算机视觉的各个领域，深度神经网络学习的特征也逐渐替代了手工设计的特征，人工智能也变得更加“智能”。

另一方面，深度神经网络的出现也降低了人工智能系统的复杂度。如图 3-14 所示，在传统的模式分类系统中，特征提取与分类是两个独立的步骤. 而深度神经网络 将二者集成在了一起。我们只需要将一张图片输入给神经网络，就可以直接得出对图片类别的预测，不再需要分步完成特征提取与分类。从这个角度来讲，深度神经网络 并不是对传统模式分类系统的颠覆，而是对传统系统的改进与增强。

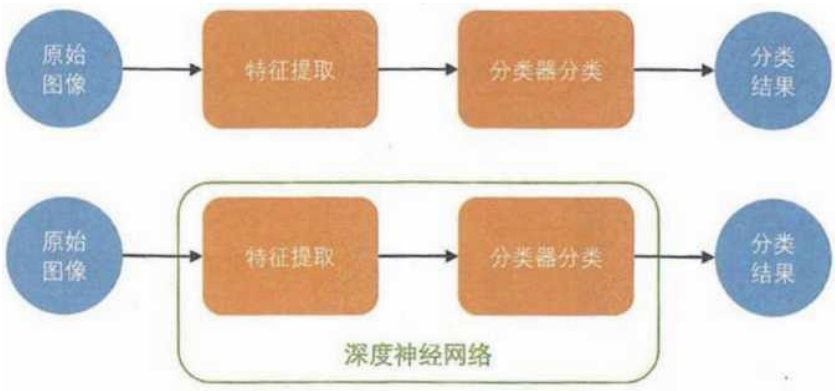


图 3-14 深度神经网络与传统 模式分类系统的区别和联系

深度神经网络的结构

一个深度神经网络通常由多个顺序连接的层（layer）组成。第一层一般以图像为输入，通过特定的运算从图像中提取特征。接下来每一层以前一层提取出的特征输入，对其进行特定形式的变换，便可以得到更复杂一些的特征。这种层次化的特征提 取过程可以累加，赋予神经网络强大的特征提取能力。经过很多层的变换之后，神经网络就可以将原始图像变换为高层次的抽象的特征。

这种由简单到复杂、由低级到高级的抽象过程可以通过生活中的例子来体会。例如，在英语学习过程中，通过字母的组合，可以得到单词；通过单词的组合，可以得到句子；通过句子的分析，可以了解语义；通过语义的分析，可以获得表达的思想或目的。而这种语义、思想等，就是更高级别的抽象。

接下来，让我们来看一个具体的神经网络的例子，以对深度神经网络的结构有一个直观的感受。这个网络中出现了卷积层、ReLU 非线性激活层、池化层、全连接层、 softmax 归一化指数层等概念，我们后面会一一进行介绍。

在图 3-15 中，我们展示了获得 2012 年 Image Net 挑战赛冠军的 Alex Net 神经网络。这个神经网络的主体部分由五个卷积层和三个全连接层组成。五个卷积层位于网络的最前端，依次对图像进行变换以提取特征。每个卷积层之后都有一个 ReLU 非线性激活层完成非线性变换。第一、二、五个卷积层之后连接有最大池化层，用以降低特征图的分辨率。经过五个卷积层以及相连的非线性激活层与池化层之后，特征图被转为 4096 维的特征向量，再经过两次全连接层和 ReLU 层的变换之后，成为最终的特征向量。再经过一个全连接层和一个 sofImax 归一化指数层之后，就得到了对图片所属类别的预测。

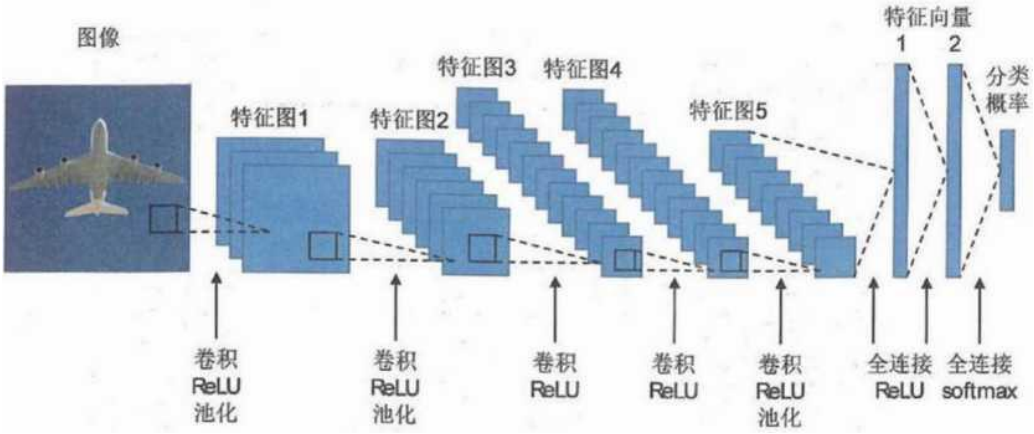


图 3-15 Alex Net 神经网络结构示意图

卷积层

卷积层(convolutional layer)是深度神经网络在处理图像时十分常用的一种层。 当一个深度神经网络以卷积层为主体时，我们也称之为卷积神经网络(convolutional neural network)。

神经网络中的卷积层就是用卷积运算对原始图像或者上一层的特征进行变换的层。在上一节中我们学习了边缘特征的提取，知道一种特定的卷积核可以对图像进行一种特定的变换，从而提取出某种特定的特征，如横向边缘或纵向边缘。在一个卷积层中，为了从图像中提取出多种形式的特征，我们通常使用多个卷积核对输入 图像进行不同的卷积操作(图 3-16)。一个卷积核可以得到一个通道为 1 的三阶张量，多个卷积核就可以得到多个通道为 1 的三阶张量结果。我们把这些结果作为不同的通道组合起来，就又可以得到一个新的三阶张量，这个三阶张量的通道数就等 于我们使用的卷积核的个数。由于每一个通道都是从原图像中提取的一种特征。我们也将这个三阶张量称为特征图(feature map)。这个特征图就是卷积层的最终输出。

特征图与彩色图像都是三阶张量，也都有若干个通道。因此卷积层不仅可以作用 于图像，也可以作用于其他层输出的特征图。通常，一个深度神经网络的第一个卷积层会以图像作为输入，而之后的卷积层会以前面的层输出的特征图为输入。

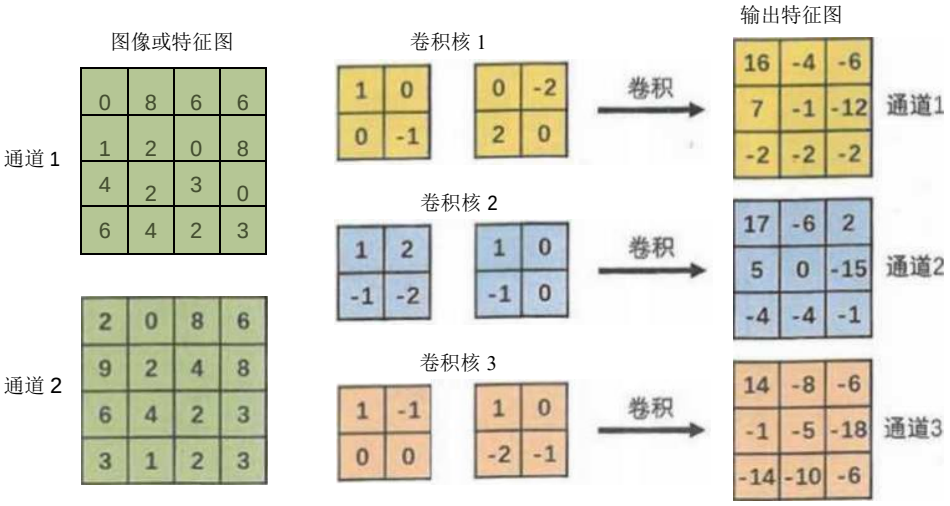


图 3-16 使用多个卷积核提取多种 特征并组合成多通道的特征

全连接层

在图片分类任务中，输入图片在经过若干卷积层之后，会将得到的特征图转换为特征向量。如果需要对这个特征向量进行变换，经常用到的便是全连接层（fully-connected layer）。

在全连接层中，我们会使用若干维数相同的向量与输入向量做内积操作，并将所有结果拼接成一个向量作为输出。具体来说，如果一个全连接层以向量 X 作为输入，我们会用总共 K 个维数相同的参数向量 W_k 与 X 做内积运算，再在每个结果上加上一个标量 b_k，即完成 $y_k = X \cdot W_k + b_k$ 的运算。最后，我们将 K 个标量结果 y_k 组成向量 Y 作为这一层的输出。

归一化指数层

归一化指数层（softmax layer）的作用就是完成多类线性分类器中的归一化指数函数的计算。具体来说，对于输入向量 $X = (x_1, x_2, \ldots, x_n)$，计算 n 个标量值 $y_k = \frac{e^{x_k}}{e^{x_1} + \ldots + e^{x_n}}$，并将它们拼接成向量 $Y = (y_1, y_2, \ldots, y_n)$ 作为输出。归一化指数层一般是分类网络的最后一层，它以一个长度和类别个数相等的特征向量作为输入（这个特征向量通常来自一个全连接层的输出），然后输出图像属于各个类别的概率。

非线性激活层

通常我们需要在每个卷积层和全连接层后面都连接一个非线性激活层(non-linear activation layer)。为什么呢？其实不管是卷积运算还是全连接层中的运算，它们都是关于自变量的一次函数，即所谓的线性函数(linear function)。线性函数有一个性质：若干线性计算的复合仍然是线性的。换句话说，如果我们只是将卷积层和全连接层直接堆叠起来，那么它们对输入图片产生的效果就可以被一个全连接层替代。这样一来，虽然我们堆叠了很多层，但每一层的变换效果实际上被合并到了一起。而如果我们在每次线性运算后，再进行一次非线性(non-linear)运算，那么每次变换的效果就可以得以保留。非线性激活层的形式有许多种，它们的基本形式是先选定某种非线性函数，然后再对输入特征图或特征向量的每一个元素应用这种非线性函数，得到输出。常用的非线性函数有：

- 逻辑函数(logistic function)(如图 3-17 左)，

$$s(x) = \frac{1}{1+e^{-x}}$$

- 双曲正切函数（hyperbolic tangent function）（如图 3－17 中），

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{-x}+e^{x}}$$

- 线性整流函数（rectified linear function）（如图 3－17 右）等。

$$\mathrm{ReLU}(x) = \begin{cases} 0, & x<0 \\ x, & x\geqslant 0 \end{cases}$$

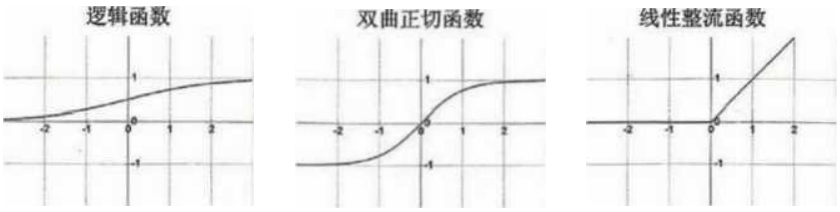


图 3-17　不同的非线性激活函数

以线性整流函数构成的非线性激活层(简称为 ReLU 层)为例，对于输入的特征向量或特征图，它会将其中小于零的元素变成零，而保持其余元素的值不变，就得到了输出。因为 ReLU 的计算非常简单，所以它的计算速度往往比其他非线性激活层快很多，加之其在实际应用中的效果也很好，因此在深度神经网络中被广泛地使用。

池化层

在计算卷积时，我们会用卷积核滑过图像或特征图的每一个像素。如果图像或特征图的分解率很大，那么卷积层的计算最就会很大。为了解决这个问题，我们通常在几个卷积层之后插入池化层(pooling layer)，以降低特征图的分辨率。

池化层的池化操作步骤如下。首先，我们将特征图按通道分开，得到若干个矩阵。对于每个矩阵，我们将其切割成若干个大小相等的正方形小块。以图 3-18 为例，我们将一个 4x4 的矩阵分割成 4 个正方形区块，每个区块的大小为 2x2。接下来，我们对每一个区块取最大值或平均值，并将结果组成一个新的矩阵。最后，我们将所有通道的结果矩阵按原顺序堆叠起来形成一个三阶张量，这个三阶张量就是池化层的输出。对每一个区块取最大值的池化层，我们称之为最大池化层(max pooling layer)，而取平均值的池化层称为平均池化层(average pooling layer)。

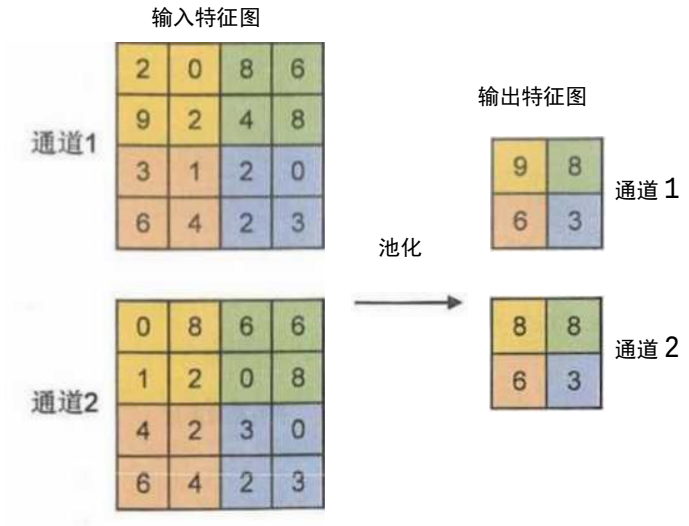


图 3-18 最大池化层示意图

在图 3-18 中，经过池化后，特征图的长和宽都会减小到原来的 1/2，特征图中 的元素数目减小到原来的 1/4。通常我们会在卷积层之后增加池化层。这样，在经过 若干卷积、池化层的组合之后，在不考虑通道数的情况下，特征图的分辨率就会远小于输入图像的分辨率，大大减小了对计算量和参数数量的需求。

人工神经网络与生物神经网络

人工神经网络最初是受到生物神经网络(biological neural networks)的启发而提出的。生物神经网络由数以亿计的神经元(nearon)相互连接而成。当我们思考或者对外界刺激做出反应时，神经元之间就在互相传递信息。人工神经元是生物神经元的一个数学模型。以人工神经元为基本单元，我们可以构建卷积层、全连接层、非线性激活层等，进而构建人工神经网络。也正是因为这样的联系，特征图或特征向量中的每个元素也称为神经元，元素的值称为神经元的响应。

但是人工神经元只是生物神经元的一个的数学模型，并不能精确描述生物神经元复杂的行为。在机器学习领域，人工神经网络的研究重点主要局限于特定的人工智能任务，在实际应用中，主流的人工神经网络和生物神经网络已没有直接的联系。

人工神经网络的训练

分类器需要经过训练才可以区分属于不同类别的特征向量，深度神经网络也需要 经过训练才能学习出有效的图像特征。我们知道，训练本质上就是寻找最佳参数的过程。在线性分类器中，参数包含所有线性函数的所有系数。而在神经网络中，卷积层 中所有卷积核的元素值、全连接层中所有内积运算的系数都是参数。为了将鸢尾花的二维向量分成两类，我们只需要训练三个参数。而在 Alex Net 中. 需要学习的参数多 达六千万个，其难度远高于线性分类器的训练。针对神经网络训练的问题，人工智能 科学家们提出了反向传播(backproppagation)算法(图 3 -19)。它是训练神经网络最 有效的手段之一。

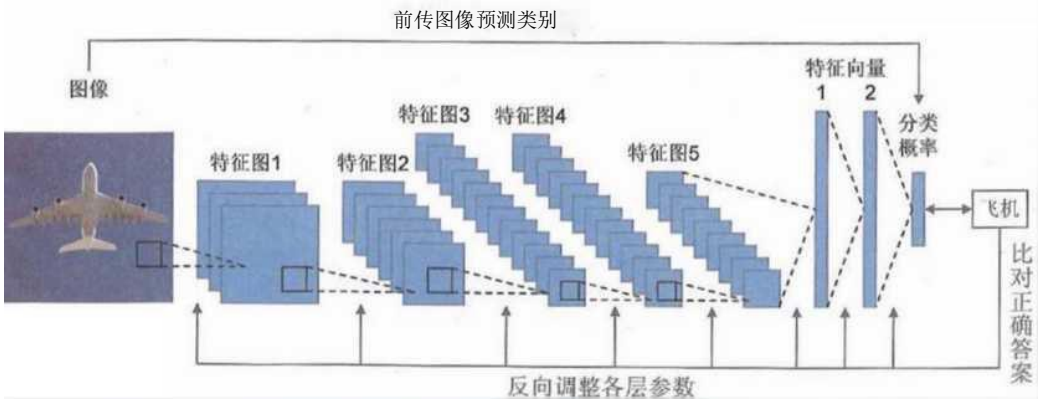


图 3-19 反向传播算法示意图

每次我们将一幅训练图像输入网络中，经过逐层的计算，最终得到预测的属于每一类的概率，我们将预测结果与正确答案进行对比，如果发现预测结果不够好，那么会从最后一层开始，逐层调整神经网络的参数，使得网络对这个训练样本能够做出更 好的预测。我们将这种从后往前调整参数的方法称为反向传播算法。具体的调整算法涉及梯度计算的链式法则(chain rule)和随机梯度下降(stochastic gradient descent) 等更复杂的知识，这里不做详细介绍。

实验 3-2：利用深度神经网络进行图片分类

在这个实验中，我们将用深度神经网络代替传统分类系统，在同样的 CIFAR 10 数据集上完成图片分类任务，我们将使用简单的 ResNet 18 神经网络完成这个任务。

1. 利用工具包中提供的函数，绘制出 ResNet 18 的网络结构图。观察并指出它与 Alex Net 的区别。

2. 利用工具包中提供的函数和 CIFAR 10 的训练集训练一个 ResNet 18 网络， 并记录在训练集上的分类正确率。

3. 利用工具包中提供的函数对第一个卷积层中的卷积核进行可视化。如果用这些卷积核对图像进行卷积运算，会对图像起到什么效果？

4. 任选一张图片输入到神经网络中，利用工具包中提供的函数，对各层输出的特征图进行可视化，体会特征图的变化过程。

5. 用工具包中提供的函数，对神经网络学习到的特征进行可视化，理解神经网络层次化特征提取的概念。

6. 利用 CIFAR 10 的测试集对训练好的 ResNet 18 网络进行测试，并记录其在测试集上的分类正确率。

7. 比较深度神经网络和传统模式分类系统的分类正确率。

3.3 “网”不厌深：深度神经网络的发展与挑战

深度之“深”

深度学习的“深”其实表征着神经网络的层数之多，更进一步代表着模型参数

之多。一个参数更多的模型，其可学习和调整的空间就更大，表达能力就更强。甚至曾经有人说过：“只要测试错误率还在下降，就可以持续不断地增加深度神经网络的层数来改进结果。”

深度神经网络模型表现的飞快提升，和网络结构不断复杂、网络层数不断增加是 分不开的。如图 3-20 所示，2012 年超越传统方法 10 个百分点的 Alex Net，共有 5 个卷 积层；而到了 2016 年的 PolyNeI，则有足足 500 多个卷积层。虽然现代网络设计并不是简单的层数的纵向堆叠，卷积层的数量并不等于网络的深度，但大体上遵循层数越多 网络越深的规律。如今在计算机视觉领域，更深的网络也屡见不鲜。这些“深”而复杂的网络，不断刷新着以往相关领域任务中的最好成绩，给我们带来一次又一次的震撼。

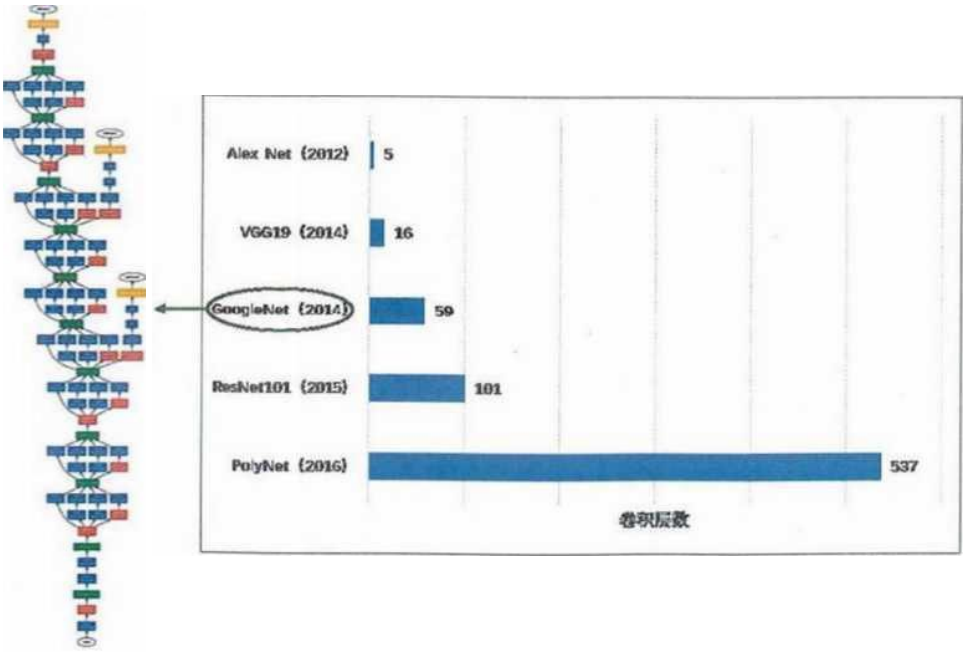


图 3-20 历年经典网络和其卷积层数

深度之“助”

正如世界上没有无缘无故的成功，深度学习的火热浪潮，不只与学科自身的积累、更新及发展有关，和一些外在推动力也有着密不可分的联系。在这里我们着重强 调两点较为重要的助力因素：数据与计算能力。

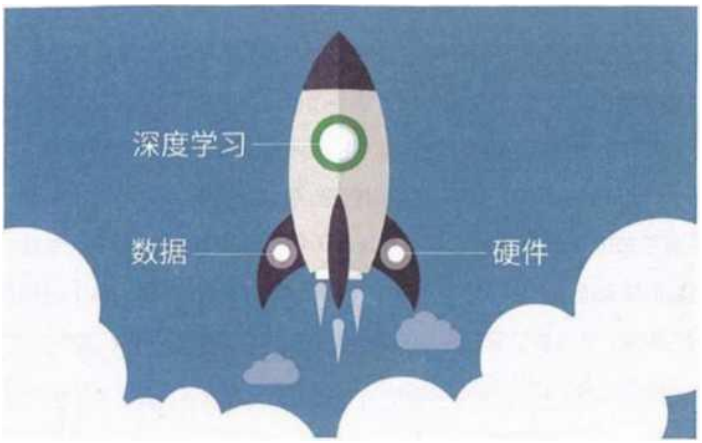


图 3-21 深度学习的助推器：数据与计算能力

数据

随着信息革命的到来，存在于互联网上的数据以指数爆炸的形式增长。大量涌现的数字资料，使得传统的数据处理方法难以对其进行分析和处理。然而这一看似棘手的挑战，对于深度学习的发展，却是绝佳的机遇。

作为一门数据驱动的科学，深度模型本身的性能就和训练数据的总量、多样性等特性有着密不可分的联系。一个“见多识广”的深度模型，对于实际问题的处理和表现往往更加“优秀”。可以说，数据就像燃料一样，推动着深度学习这枚火箭不断前进。

计算能力

虽然人工智能科学家们对于深度神经网络的构想十分精妙，但是构想要落地也需要硬件资源的支持。具体地说，神经网络的训练过程需要大量的计算资源，而越深层越复杂的网络对计算资源的需求就越大。

这种繁重的计算任务是普通 CPU 难以负荷的，与此同时，更强大的图形处理器 (GPU)的出现和不断更新一定程度上成就了当今深度学习大热的局面。以深度神经网络中的“先驱者”——Alex Net 为例，为了完成 Image Net 分类模型的训练，使用一颗 16 核 CPU 需要一个多月才能完成，而使用一块新型的 GPU 则只需要两三天，大大提高了训练效率(如图 3-22 所示）。

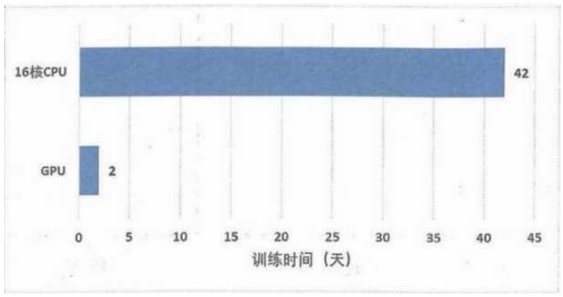


图 3-22 在不同硬件设备上训练 Alex Net 的时间对比

深度之“难”

如果仅仅通过不新加深网络，我们就能得到性能更好的模型，那么深度学习领域的一切研究是不是就变得十分容易、只要通过不断加深网络便可解决所有问题？然而真实情况并不是如我们所想的这样简单。事实上，更深的神经网络除了会带来更加巨大、令人难以负担的资源消耗外，其在对应任务上的表现有时却会不升反降。这种现象到底为什么会发生呢？下面我们将为大家介绍两个重要原因。

过犹不及：过拟合和欠拟合

过多的层数带来过多的参数，很容易导致机器学习中一个常见的通病：过拟合。训练模型的过程是在训练集上完成的，而我们对一个模型表现的评测会在测试集上完成。有的模型在训练集上是一等一的“优等生”，但是在测试集上的表现却不尽如人意，有时的表现都不能达到及格水平。我们将这种复杂模型过多地“迎合”训练数据、导致其在大量新数据上表现很差的现象称为过拟合(overfitting)。就好比在学习解数学题时，有时候会遇到一些偏题、怪题，甚至参考答案错误的题目，但我们却喜欢钻牛角尖，死记硬背这些题目的解法，然后机械地套用到正常题目上而导致做错。而欠拟合的模型则是一个反应有些愚钝的“差生”：在训练数据和新数据上的表现都不能让人满意，简而言之就是能力有限。这种由于模型本身过于简单能力较弱，而导致的在训练过程中准确率很低并且难以提升、在新数据上表现同样很差的现象称为欠拟合(under-fitting)。

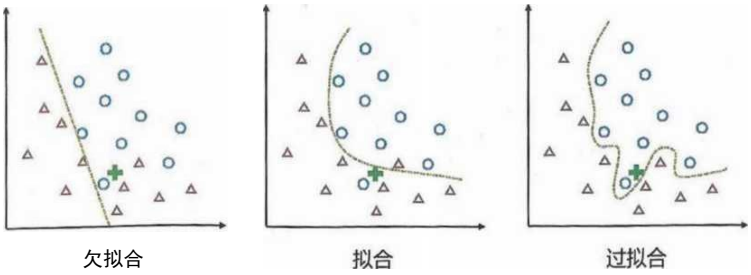


图 3-23 欠拟合、拟合、 过拟合示意图

在图像分类任务中，过拟合可以理解为由于模型的能力太强大，在训练过程中不但记住了训练集中需要被分类样本的真实特征，也记住了训练集中的很多噪声信息。这导致模型虽然在训练集上分类准确率可以达到很高甚至 100%，但在对新加入的数据进行分类时，这个死板的“优等模型”所给出的结果往往是错的，如图 3-23 所示，新来的数据(绿色十字)本来应该是红色三角的一类，却被错分为蓝色圆圈的一类。而欠拟合的模型则因为参数太少，无法完全提取和刻画待分类对象的全部特征，所以其只能将数据“草草”分开，分类准确率相对较低。

那么如何在加深网络结构、增加网络表达能力的同时，尽量避免过拟合的情况呢？神经网络训练中常用权值衰减(weight decay)等正则化(regularization)方法来解决这个问题。相应的技术细节在这里不一一展开，大家有兴趣可自行查阅相关资料。

远水难解近渴：梯度消失

我们在一定程度上解决了网络加深带来的过拟合问题后，却发现简单的层数堆叠仍然使得深度模型解决问题的能力不升反降。既然这种能力的降低与过拟合无关，那么又是什么原因呢？

事实上，对网络进行简单堆叠加深还会导致一种影响性能的现象：梯度消失 (gradient vanish)。那么什么是梯度，梯度又为什么会消失呢？

在第二章我们简单介绍了优化的概念，优化的目标就是使得网络输出的预测值和 目标值更加接近。深度网络的训练也是一种优化的过程--在这里，寻找一个在特定任务上表现较好的网络。如果将网络训练的过程比喻成下山，优化过程就相当于在连绵起伏的山脉中找到最低的那个低谷。而梯度，就相当于在行走的每一步中对优化方

向的指引（如图 3-24 所示）。

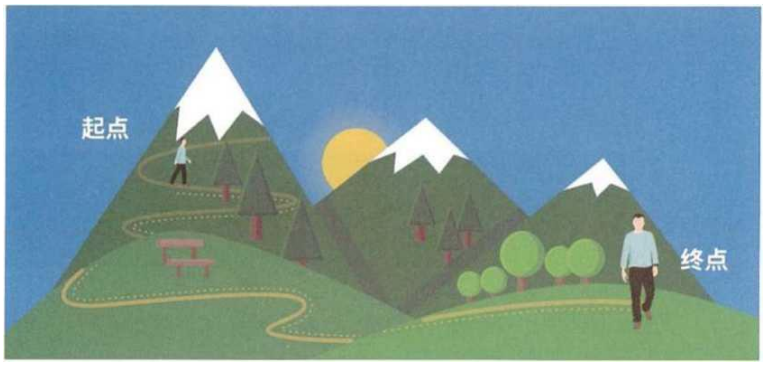


图 3-24 下山示意图

这种“指引”最直接的来源就是模型的输出结果与目标输出的差别，也就是误 差。在前面提到的反向传播算法中，我们对每一层参数的调整正是通过这种误差的传 播完成的。如果网络过深，这种来自遥远的输出层的误差会逐层地被指数级缩小（或 放大）。假设每经过一层，数值范围变成原来的一半，那么经过 10 层后，$0.5^{10}=0.00097$，意味着倒数第 11 层的梯度数值是最后一层梯度的千分之一。此时的“梯 度”将逐渐接近于 0，也就是说梯度在反向传播的过程中逐渐“消失了”。梯度消失 后，网络的优化过程失去了指导，将无法找到一个较好的解（如图 3-25 所示）。

图 3-25 迷茫的网络

解决这个何题常用的技术有批处理化(batch normalizalion)和跨层连接(short-cut)等，我们不详细探讨，有兴趣的同学可以查阅相关资料。

3.4 忽如一夜春风来：图像分类在日常 生活中的应用

学习了这么多深度学习和图像分类的知识，除了可以帮铭铭整理相册之外，我们还可以做些什么呢？其实图像分类技术在日常生活中已经处处可见，有着广泛的应用，比如人脸识别、图像搜索等等。我们以人脸识别为例，介绍一下图像分类和深度学习给生活带来的变化。

2014 年，香港中文大学团队的工作使得机器在人脸识别任务上的表现第一次超越了人类。从这一里程碑式的事件开始，“人脸识别”也成为深度学习算法着力研究 的任务之一，并在不断的发展和演进中变成了最先实现落地和改变我们生活的深度学习应用之一。

在深度神经网络被应用于“人脸识别”任务之前，传统的机器学习算法也曾试图解决这一问题。但是由于传统算法在进行识别的过程中，无注同时确保准确率与识别效率，这一情况使得传统人脸识别算法很难达到应用规模。而当前的在亿万级别人脸数据上训练得到的深度模型，在使用时则可以同时满足大规模和高精度的要求，真正应用于生活的方方面面。

人脸识别是从一张数字图像或一帧视频中，由“找到人脸”到“认出人脸”的过程，其中“认出人脸”就是一个图像分类的任务。具体地，整个识别过程一 般包括以下几个步骤：人脸检测、特征提取、人脸比对和数据保存与分析。人脸检测即对包含用户脸部的图像进行检测，找到人脸所在的位置、人脸角度等信息，也就是完成“看得到”的过程。特征提取则是要让机器“看得懂”：通过对 人脸检测步骤中检测出的人脸部分进行分析，得到人脸相应的特征，如五官特 点、是否微笑、是否戴眼镜等特征信息。这两步得到的信息，将被用于与人脸数 据库中已经记录的人像(如身份证照片)以一定的方法相比对，也就是解决“跟谁像”的问题。最后，这些分析结果将根据具体的情况被使用，服务于最终的实际应用场景。

让机器“看得到”、“看得懂”、“认得出”人脸这件事，本身有什么意义呢？下 面我们通过几个具体的应用示例，带领同学们一起感受一下，作为深度学习典型代表 的人脸识别技术，是如何为我们的生活带来美好变化的。

刷脸时代：人脸识别让生活更便捷

从第一台计算机诞生至今，我们的生活经历了由互联网产业发展所带来的巨变，足不出户可知天下事，开始进入“信息时代”。而自杰弗里•辛顿等人于 2006 年提出 深度学习的概念，这个可以使得神经网络具有更高效、更强大能力的技术热点，在短 暂沉寂之后，以一种不可阻挡的态势迅速改变着我们的生活，带领我们进入“人工智 能时代”（如图 3-26 所示）。

图 3-26 丰富的“刷脸”应用场景

以人脸识别技术为切入点和代表，我们可以实实在在地感受到我们的生活正在便捷化出门购物时，我们不用再一次又一次地输入密码，只需要“刷脸”即可迅速完成支付，新颖又便捷；地铁出行时，我们不用担心忘带地铁卡，进站时进站系统将自动进行人 脸扫描，“届嘘”即可进站，这样的方式也很好地避免了排队买票和排队进站的情况，极大地节约了出行时间；早上上班或进入学校，再也不需要考勤人员辛辛苦苦进行记录，门口的刷脸考勤系统将会自动识和记录你进入的时间，保证了考勤记录的真实公正。还有刷脸解锁、刷脸取款等等，人脸识别技术的具体应用不胜枚举，我们的生活也在强有力的人工智能技术的推动下，逐渐进入更丰富、更便捷、更美好的新纪元。

天网恢恢：人脸识别技术助力安防

车站里人山人海，偷窃国宝的嫌疑人就混在人群之中。然而人海茫茫，便衣民警们搜寻嫌疑人的难度无异于大海捞针。此时，远在城市另一处的指挥部大厅里，高清监控的影像被投影在屏幕上，监控视频中黑压压的人群如潮水般涌动，人眼几乎无法从监控视

频中捕捉到有效信息。但是民警们并不苦恼于此，因为人工智能技术早已帮他们完成了这 一切。人潮中每一张人脸都被精确地检测和识别，近乎实时地完成了特征提取与分析，并与罪犯中库的通缉犯行比对。突然，警报声响彻指挥大厅，嫌疑人被找到了！“火眼金睛”的人脸识技术已经精确地在人群中找到并框出了嫌疑人的位置，指挥员迅速向现场 便衣民警们下达指令，不出片刻，犯罪嫌疑人就已经落网。

图 3-27 人脸布控系统界面示意图

上面描述的就是人工智能中人脸识别技术助力公安抓捕犯罪嫌疑人的场景。随着人口数量的增加和人口流动性的增强，安保工作的重要性逐渐增加，并面临更大的挑战。触犯法律的人总是存在一些侥幸心理，认为自己可以隐姓埋名或匿于市井，逃避法律的制裁。而完善的人脸布控系统将让所有的罪犯都无从藏匿。人脸布控系统搭载前沿的人脸识别技术，配合逐渐完善的监控网络，通过构建一套从感知、预警、分析到决策的自动化系统，能从监控视频中实时提取有用信息，让我们的监控系统从“看得清”逐渐升级到更高层次的“看得懂”。

相信终有一日，所有罪犯都将在严密的人脸布控系统中无处遁形，那么我们的社会将与《礼记•礼运》大同篇中所描绘的美好愿景——“是故谋闭而不兴，盗窃乱 贼而不作”一越来越接近。

3.5 本章小结

在本章中，我们学习了如何进行图像分类。首先我们了解了图像在计算机中的表示，知道了图像特征提取是对三阶张量进行特定的数学运算。在手工特征提取中，我

们学习了卷积运算，并了解如何利用卷积运算提取图像特征，例如边缘特征和方向梯 度直方图。手工设计特征有其局限的一面，而用深度神经网络自动学习特征已经是图像处理中广泛应用的方法，所以我们重点学习了如何用深度神经网络来进行图像分类。我们对深度神经网络的结构有了初步的认识，了解了深度神经网络中一些基本的层，例如卷积层、池化层、全连接层、非线性激活层和归一化指数层。同时我们知道了利用反向传播算法进行网络训练的过程。最后，我们认识了深度神经网络的发展和挑战，也了解了图像分类在日常生活中的广泛应用。

通过本章知识的学习和相关实验的完成我们可以体会到，相比经典方法，深度学习具有更为强大的表达力，因此能更好地完成复杂的任务，与此同时，多层神经网络的训练也需要更多的数据与计算能力的支持。

从萌芽到冷遇再到复兴，深度学习的发展也经历了曲折，并不断遇到新的问题和挑战。与此同时，这些挑战是对当前研究的一种激励和推动，推动着现有研究成果的 不断完善，也推动着新技术的不断发展。石落浪起，热潮未已，深度学习将往哪里去，可以走到多远的远方，能为我们心目中的美好明天带来多少贡献，一切还有赖于在座的各位，有赖于即将投身这一领域的新生力量。

第四章 耳听八方：析音赏乐

周末铭铭去听了两场音乐会。第一场是摇滚乐队的演唱会，在有力而连贯的节奏中，或激昂或忧伤的旋律伴随歌手竭力的演唱直击内心。第 二场是管弦乐音乐会，各种乐器的声音交织在一起，产生了丰富的变化.此刻像在聚会上欢饮畅谈，转眼又似在战场上短兵相接。回到家，铭铭思考着声音竟有这么多美妙的变化，计算机能不能像人一样欣赏这些动听的旋律呢？……

《史记•刺客列传》中记载着一句悲壮的歌词：“风萧萧兮易水寒，壮士一去兮不复还。”在高渐离击筑的伴奏下，荆轲唱出了自己视死如归的英雄气概。

《诗经•关雎》中亦有描述：“窈窕淑女，钟鼓乐之。”动情的钟鼓之声诉说的是对贤淑女子的爱慕。

在世间的千千万万种声音里，有两种声音与人的关系尤为紧密，一种是人类发声器官发出的具有信息交流意义的声音，也称语音(speech);另一种是人类创造的有 节奏和旋律的声音艺术，即音乐(music)。语音在日常交流中扮演着重要角色，我们 每天要与很多人打交道，向老师请教问题，和同学讨论时事，与父母分享趣闻，等等，语音无疑是最高效的工具。有趣的是，人的听觉对语音也有偏好，在嘈杂的会场上,人能够准确辨识出与自己交谈的人所说的话，这也被称为鸡尾酒会效应(cock- tail party effect)。音乐是一种独特的艺术，当跃动的旋律与情绪相碰撞时，我们的思 绪就全部被音符占据，陶醉于其中，感受到灵魂的洗礼，这是多么美妙啊！音乐的历史源远流长，欢快的钟鼓乐和悲歌有着完全不同的风格，就如同当今有爵士、摇滚等 不同的曲风，歌手有各自擅长的风格，听众有自己喜好的类型，这即是音乐领域的百家争鸣。

在上一章里我们给计算机装上了一双“千里眼”，让它学会了识别图像；这一章里我们要借助人工智能，赋予计算机一对“顺风耳”，让计算机也能听懂语音，欣赏音乐，探索声音的奥妙。

4.1　洗耳恭听：听声的艺术

人耳听声

我们在物理课上就学习过声波的产生和传播原理，声波由物体振动产生，经介质传播，最后到达人耳被人感知。我们在生物课上也学习过人耳结构，如图 4-1 所示，声波由耳郭收集之后经一系列结构的传导到达耳蜗，耳蜗内有丰富的听觉感受器，可将声音传导到听神经，最后引起听觉。

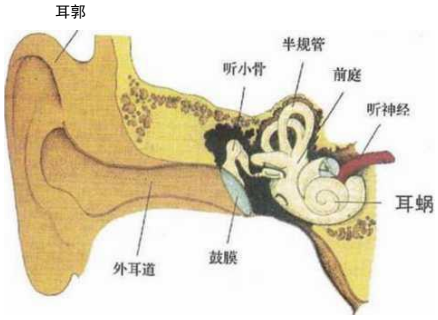


图 4-1　人耳结构

频率（frequency）是声音的重要特征，它代表了发声物体在一秒内振动的次数，单位是赫兹（Hz)。人耳的精妙结构也决定了我们对不同频率的声音有着不同的敏感程度，如图 4-2 所示，横坐标代表频率，纵坐标代表引起人耳听觉的声音强度，单位是分贝（dB)，这个值越小代表人对该频率的声音越敏感。有趣的是，人耳最敏的频率与婴儿发声的频率大致相同。人发声的频率范围在 85-1100Hz, 可以看到人耳对这段范围内的频率也是相对敏感的，这就是说人耳的构造很大程度上为人与人之间的交流提供了便利，试想如果我们连超声波都听得一清二楚，那世界会变得多么喧嚣。

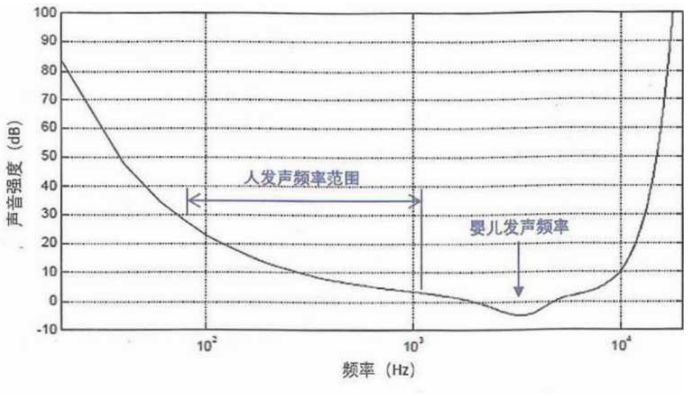


图 4-2 人耳听觉对不同频率声音的敏感程度

声音的数字化

计算机没有耳朵，那它怎么感知声音呢？这时候就需要把声波转换为便于计算机存储和处理的音频文件了（如 MP3 格式）。这个过程如图 4-3 所示，从声波到最终的 MP3 文件主要经历了采样（sampling）、量化（quantization）和编码（encoding）等步骤。

首先我们通过话筒中的传感器把声波转化为电信号（如电压），这就好比耳蜗中的听觉感受器把声波传导到听神经。但是计算机是无法存储连续信号的，因此我们需要通过采样使得电信号在时间上变得离散，再通过量化使得它在幅度上变得离散。声音变成了离散的数据点，计算机就可以通过不同的编码方式将它存储为不同的文件格 式，我们听音乐时常用的 MP3 就是其中一种。计算机里面的音频文件描述的实际上是一系列按时间先后顺序排列的数据点，所以也被称为时间序列（time series），把它可视化出来就是我们常见的波形（waveform），其横坐标代表时间，纵坐标没有直接的物理意义，它反映了传感器在传导声音时的振动位移。因为振动位移随时间在 0 附近反复振荡，因而波形也是随时间在 0 附近不断振荡的。当采样频率比较高时，波形看起来是近似连续的。

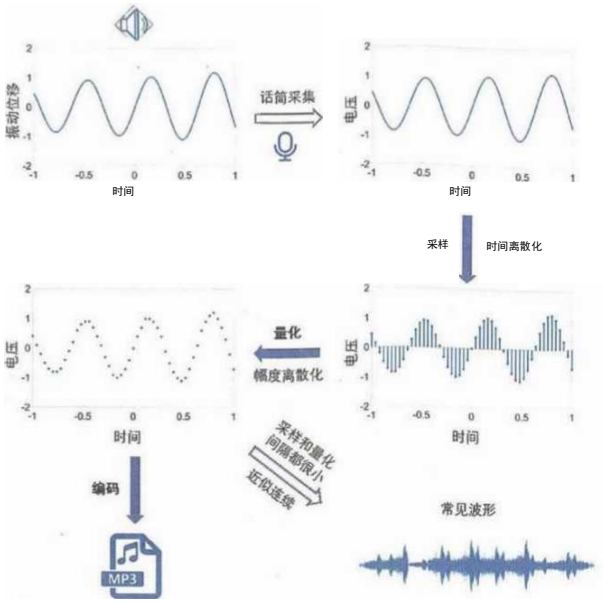


图 4-3 声音的数字化

知识链接：采样频率

采样频率（sampling rate) 也称采样率，采样率越高，声音还原得越自然。通常 MP3 格式的采样率是 44100Hz 因为人耳对高频的声音不敏感，因此继续增加采样率对于听觉感受的影响很小，却会浪费存储空间。

与图像类似，声音数字化后的取值范围也是有限的。常见的音频一般有两个声道（对应左耳、右耳），而图像通常有三个通道（对应红、绿、蓝）。在本章中若不特别说明我们只考虑一个声道。

实验 4-1：观察声音的波形，理解声音的数字化

1. GTZAN 是一个包含了不同风格音乐的数据集，从该数据集中选取一个你喜欢的音乐片段，画出该音乐片段所对应的波形。
2. 从不同的时间尺度观察波形，思考听觉感受与波形的视觉呈现之间的联系。

通过频谱理解乐音三要素

通过声音的数字化，计算机“听”到了声音，那么计算机如何“理解”声音呢？ 这里我们以乐音三要素为例学习一个计算机分析音频的常用方法——频谱（frequency spectrum）。如图 4-4 所示为一段音乐的波形（左）和频谱（右）。

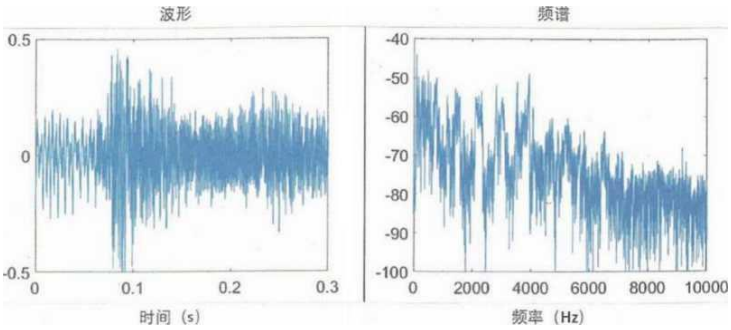


图 4-4 音乐片段的波形与频谱

频谱的横坐标代表频率，纵坐标表示频谱幅度，其含义是相应频率的声音所对应的振幅。因为一段音频中不同频率的声音强度相差很大，所以频谱幅度通常使用对数坐标，即振幅每相差 10 倍，频谱幅度相差 20 倍。例如频谱上 1000Hz 对应的幅度为-50，5000Hz 对应的幅度是-70，那么说明 1000Hz 的声音振幅比 5000Hz 的声音振幅大 10 倍。频谱图反映了不同频率的声音所占能量的多少，而我们通常只关注谱幅度的相对大小。比如一段 合唱中高音强低音弱，那么在一定范围内频率高的区域对应的频谱幅度就大，反之则频率 低的区域对应的频谱幅度大。

我们在初中物理课上接触过乐音三要素的概念，即响度、音调、音色，它们可以描述乐音的特性。

响度：最直观的乐音要素，代表产音的强弱，可由波形的幅度表示。

音调：表示人听到的声音调子的高低，声音的频率较高，听起来调子就高；声音的频率较低，听起来调子也就较低。因此可以用频谱来描述音调。

音色：是一种更复杂的特征，即便是相同的音调和响度，用不同的乐器演奏或者不同的人来演唱都有不同的听觉效果。这是因为乐器和声带在振动发声的过程中，除了音调所对应的频率 f 以外，还伴有一些高频成分（频率 2f，3f，4f...），也称泛音。

这些高频成分对应的幅度各有不同，于是造就了独特的听觉感受。

我们用吉他和钢琴举个例子。如图 4-5 所示，左侧是拨动一根吉他弦发出的声音波形和其频谱，右侧是钢琴某个按键发出的声音波形和它的频谱。从二者的波形图中很容易看出响度由大变小，从频谱图中可以发现有一系列的峰值，其中第一个最高峰所处的频率即为音调，而在这个频率的整数倍的位置都有不同大小的峰值，它们之间的比例不同就反映出声音音色的不同。

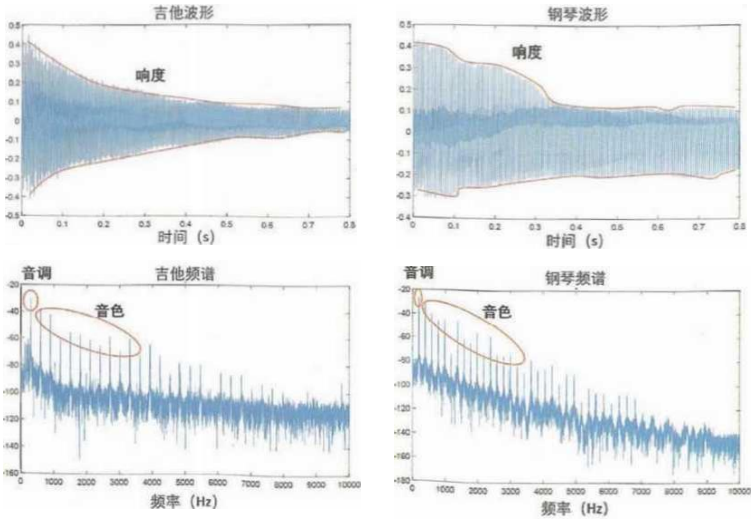


图 4-5 吉他和钢琴的波形与频谱

实验 4-2：用频谱图分析乐音的特点

1. 现有两个音频，它们分别是钢琴和吉他演奏的一个单音。画出它们的频谱图，理解频谱图横、纵坐标的含义。

2. 通过频谱图确定它们演奏的音调，并说明二者音色的差异。

4.2 丝竹管弦：音乐风格分类

在这一节里，我们要让计算机更深入地理解声音，并完成一个音乐风格分类的任 务。音乐风格是由各种音乐要素富有个性地结合在一起而产生的。我们的任务是让计 算机“听”一段 30 秒长的乐曲，然后判断其所属的风格类型（genre）。在这个任务中 一共有 10 种可能的音乐风格，包括爵士（jazz）、摇滚（rock）、嘻哈(hip-hop）等。

•输入：一段 30 秒长的乐曲所对应的时间序列

•输出：音乐风格（10 选 1）

计算机“耳中”的风格

人可以凭借感官和经验来判断音乐风格，计算机也需要在大量“听”音乐的过程中形成它的“经验”，然后依据“经验”，对音乐进行分类。因此，数据对于计算机 而言尤其重要。没有各种类型的音乐数据,再强大的人工智能算法也难以发挥作用，这就是“巧妇难为无米之炊”

乐音数据如何变为计算机的“经验”呢？回顾第三章里我们已经完成的图像识 别的例子我们先从图像数据中提取特征,然后再用分类器对特征进行分类。音乐风 格分类可以如法炮制，如图 4-6 所示，我们把任务拆成两部分。第一，设计一个特 征提取器从音乐中提取特征；第二，训练一个分类器根据音乐特征判断其风格类型。

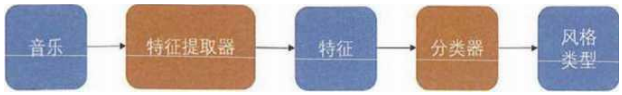


图 4-6 音乐风格分类流程图

回顾与新知：特征

特征通常是一个比数据短得多的序列，但包含了数据中最有代表性的信息。比如我们可以对一段音乐做出这样的描述：“鼓、吉他和贝斯的声音交织在一起，节奏强弱有节律地变化，让人情不自禁地跟着一起摇摆。”根据这样的描述不难想象这段音乐是一段摇滚，所以鼓、吉他和贝斯以及引起听众的摇摆可以作为摇滚音乐的“特征”，这种特征便于人的理解，但只能用语言来进行表达和接受。若要设计便于计算机理解的特征，就必须进入乐曲的数据中寻找。

我们可以把音乐的时间序列看作一个向量，对于一段时长 30 秒的音乐，采样频率为 44100Hz，这段音乐对应的向量的维数是多少呢？

由采样频率知道每秒钟的时间序列可以表示为 44100 维向量，那么整段音乐的维数是 30 x 44100 = 1323000，大约是 130 万。一张 1000 x 1000 分辨率的黑白图片包含了 100 万个像素点，表示成向量也有 100 万维，这和 30 秒长的音乐差不多。使用分类器直接对这样高维数的数据进行分类在实际中效果很差，而且给训练分类器带来很大的计算负担。所以从音乐数据中提取好的特征是非常重要的一个环节，我们将在下一小节介绍一个经典的特征。

经典的声学特征：梅尔频率倒谱系数

我们已经学习了频谱的概念，它可以直观地反映出乐音三要素的信息。但频谱的数据维数和音乐的数据维数是相同的，直接使用频谱进行分类依然困难，它并不是一个很好的特征。这里我们要学习一个比频谱更加有效且被广泛使用的特征——梅尔频 率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)。MFCC 特征的维数很低，它可以粗略地刻画出频谱的形状，因而可以大致描述出不同频率声音的能最高低。不仅如此，这种对频谱的粗略刻画还可以表达出声音的一个重要特性一共振峰。

拓展阅读：共振峰

共振峰(fommnt)指的是声音频谱上能址相对集中的一些区域。共振峰在语音的分析中较为常用，因为它在元音（vowel)的频谱上十分明显，而且不同元音的共振峰也有显著的区别。比如在图 4-7 中，元音。（左）和 i（右）在频谱图上的共振峰位置就明显不同。

频谱上为什么会有共振峰呢？原来，人说话的时候口、鼻、咽等形成了一个连通的腔体，特定频率的声音能在这个腔体里产生共振而得到放大，在频谱图上就表现为共振峰。乐器的发声过程也可以和人的发音相类比，比如提琴的琴身是一个共鸣箱，它的作用类似于人的口腔，某些频率的声音可以在共鸣箱中产生共振，形成独特的音色。

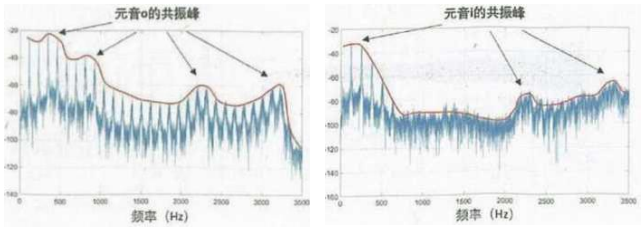


图 4-7 语音频谱的共振峰

既然 MFCC 特征有这么多优点，那它是如何提取的呢？正如它的名字，我们要先用梅尔频率对频谱进行处理得到一组 26 维的特征，然后再计算它的倒谱得到最终的 13 维 MFCC 特征。下面我们来学习一下这两个步骤的具体过程。

如图 4-8 所示，梅尔频率（Mel-Frequency）是一种特殊的频率刻度，它与普通频率的函数关系为 $mel(f) = 1125 \times \ln(1 + f/700)$。梅尔频率刻度下等长的频率区间 对应到普通频率下变为不等长的区间：在低频部分分辨率高，高频部分分辨率低。这与人耳的听觉感受是相似的，即在一定频率范围内人对低频声音比较敏感而对高频声音不敏感。在每一个频率区间对频谱求均值，它代表了每个频率范围内声音能量的大小。一共有 26 个频率范围，从而得到 26 维的特征。

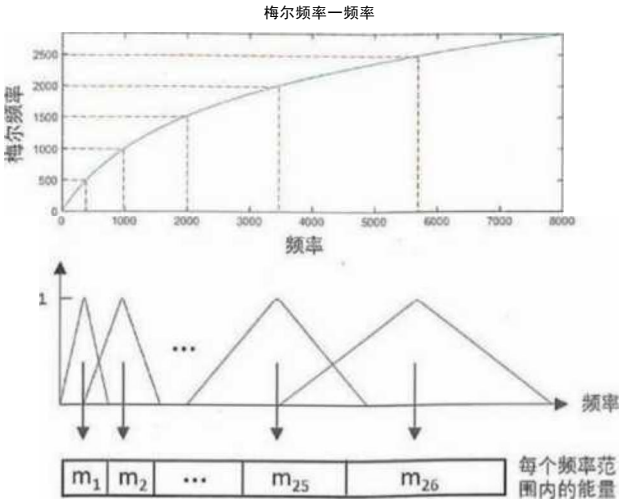


图 4-8 梅尔频率

倒谱（Cepstral）是由上述 26 维特征再做数学变换后得到的，进一步把特征维数降低到 13 维，这样我们就得到了 MFCC 特征。具体的变换过程较为复杂，需要了解的是，这 13 维特征仍然反映了音频信号在不同频率范围内的能量大小。其中保留了音频信号的一些重要特点，包括我们所学过的共振峰。

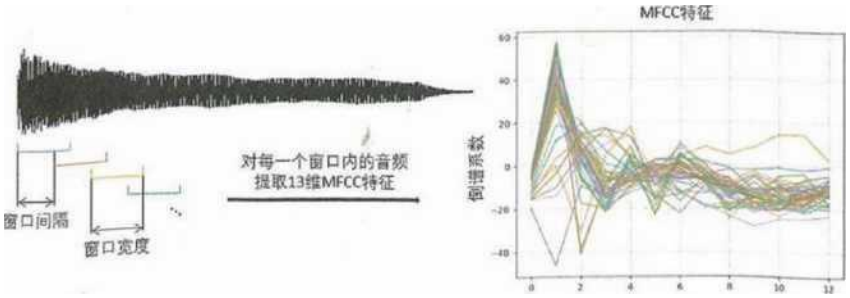


图 4-9 提取 MFCC 特征

图 4-9 展示了提取 MFCC 特征的一个示例。先把音频切分为等间隔的若干小段（可重叠），然后对每一小段分别提取 13 维的 MFCC 特征。在切分音频的时候有窗口 宽度和窗口间隔两个参数，这些参数可以根据音频的特点进行调节。一种典型的参数 是窗口宽度 25 毫秒，窗口间隔 10 毫秒。

实验 4-3：观察并理解 MFCC 特征

1. 选取 GTZAN 数据集中的任一段音乐，从中截取若干长度约 25 毫秒的片段，保证有些片段在时间上相邻，有些在时间上间隔较远。

2. 画出与这些音乐片段对应的频谱及 MFCC 特征，观察这些频谱和 MFCC 特征的异同。

深度学习方法

回顾一下我们的任务，其核心有两部分：第一是提取特征，第二是对特征分类。 我们已经介绍了 MFCC 特征，现在可以设计一个分类器对音乐的 MFCC 特征进行分类。 为了达到较高的准确率，我们使用神经网络完成分类任务。其输人是音乐的 MFCC 特征，输出是其风格类型。实际上，这个神经网络是在 MFCC 的基础上提取了更加强大的特征，并用这个特征完成了风格分类。

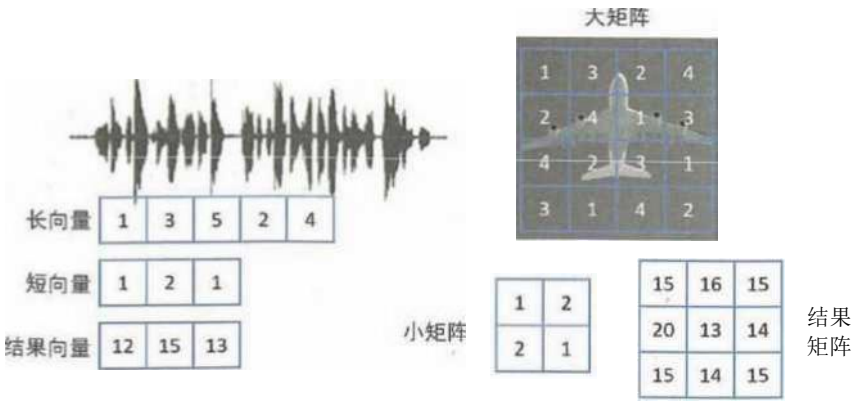


图 4-10 一维卷积 vs 二维卷

在第三章中我们学习了卷积层和池化层，并用它们提取了图像的特征。我们可以用相似的结构来提取音频的特征，但不同的是，这里的卷积和池化是作用在音频上的。音频只有一个时间维数，而图像有两个空间维数，所以提取音频特征的卷积核与处理图像的不同，如图 4-10 所示。快来验算一下图中的卷积结果是否正确吧。

经过了卷积层和池化层，神经网络提取了比 MFCC 更加强大的特征，接下来我 们要对这些特征进行分类。和图像识别的任务类似，我们首先通过全连接层得到一个长度与风格类型数相同的序列，然后用归一化指数层得到音乐属于每一种风格的概率。

实验 4 -4 ：使用 MFCC 特征和神经网络完成音乐风格分类

1．准备训练和测试数据，提取训练和测试音乐的 MFCC 特征。

2. 使用结构形如“卷积一池化一全连接一 softmax”的神经网络（参数可自行设计）进行特征的提炼和分类，其输入为音乐的 MFCC 特征，输出为音乐的风格类型。训练这个神经网络。

3. 测试音乐风格分类的准确率。

4.3 言听计从：语音识别技术

语音识别的应用

语音识别(speech recognition)的目的是把人说的话转化为文字或者机器可以理解的指令，从而实现人与机器的语音交流。语音识别技术已经在现实生活中得到了广泛的应用。还记得我们的主人公铭铭吗？铭铭有写日记的习惯，但从上高中开始，他就不再使用日记本了，而是直接用语音输入法将他一天的精彩生活口述录入到手机里，十分方便。利用语音识别技术，机器成为一位合格的笔录员。除此以外，机器还能理解人讲的话，现在很多智能手机都提供了语音助手。平时发微信铭铭都很少打字了，他直接对语音助手说“给爸爸发条微信”，然后说出发送的内容，一条微信就发了过去。发短信，打电话，叫出租车，这些日常的事情都可以通过对话的方式轻松实现。可以想象，在十年后的未来，铭铭将拥有一个家政机器人，它不仅可以听懂语音 指令完成家务，还能参与家庭会议为全家旅行出谋划策；铭铭的医生朋友也会有一个智能的机器人助理，它可以听口述记录病例，根据语音指令调取检查结果，甚至加入到治疗方案的讨论。语音识别技术将在更大程度上为人类提供便利（如图 4- 11 所示）。

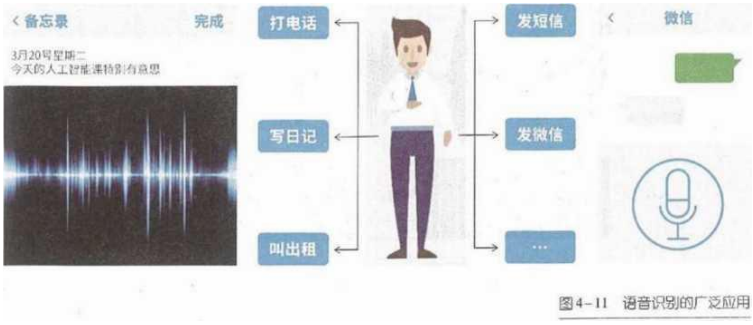


图4-11　语音识别的广泛应用

语音识别的原理

语音识别是一个非常复杂的任务，想要达到实用的水准并不容易。我们也可以把语音识别理解成一个分类任务，即把人说的每一个音都找到一个文字对应。然而这个分类任务却比我们刚刚完成的音乐风格分类复杂得多。音乐风格分类只需要对一整段音频做一次分类，而且其类型数目较少；语音识别需要对每一个音都进行分类，文字　的数量成千上万，可能的类别数也很多。可以想象.这样的分类任务是非常困难的。 但是语音识别也有它简单的一面，人类的语言是很有规律的，我们在做语音识别的时　候应该要考虑这些规律。第一，每种语言在声音上都有一定的特点，以汉语为例，我们都学过拼音，不认识的字我们通过拼音就能知道它的发音了。拼音的声母和韵母的 数量比汉字的数量少很多,我们可以用汉语的声学特性提高语音识别的准确率。第 二，汉语的语言表达也有一定的规律，比如我们根据声音的特性识别出来一个词 “hao chi”，那么这个词更有可能是“好吃”而不是“郝吃”，因为前者在汉语的表达中具有一定的意义而且会经常出现。

图 4-12 为语音识别流程图。首先把一段语音分成若干小段，这个过程称为分帧。然后把每一帧识别为一个状态，再把状态组合成音素，音素一般就是我们熟知的声母和韵母，而状态则是比音素更加细节的语音单位，一个音素通常会包含三个状态。把一系列语音帧转换为若干音素的过程利用了语言的声学特性，因而这一部分被称为声学模型(acoustic model)。从音素到文字的过程需要用到语言表达的特点，这样才能从同音字中挑选出正确的文字，组成意义明确的语句，这部分被称为语言模型 (launguage model) 。

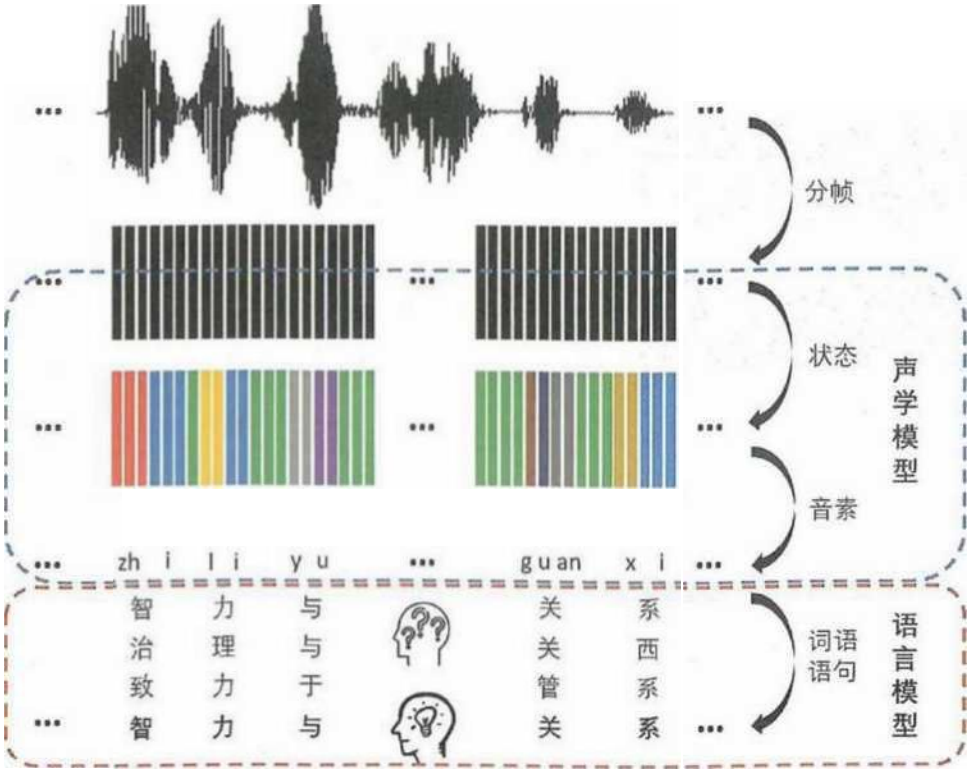


图 4-12 语音识别流程

思考与讨论：语音识别的准确率与什么有关？

语音识别的准确率与声学模型及语言模型都密切相关。例如一个语音识别系统，其声学模型可以描述普通话的发音特点，语言模型可以描述常用话题的语言表述。如果用这个语音识别系统识别播音员播报新闻的语音， 那么准确率就很高。但是用它识别一个带有口音的老师在讲述文言文阅读时的语音，其准确率就会降低很多。

4.4 听声辨曲：乐曲检索技术

在一些音乐应用中有一个有趣的功能，根据用户哼唱的一个片段找出相对应的歌曲，这就是乐曲检索。乐曲检索任务的输入通常是一个很短的音乐片段，而输出是数据库中与输入片段最为相似的乐曲。在这一节里我们将学习一种可以实现乐曲检索的简单方法。

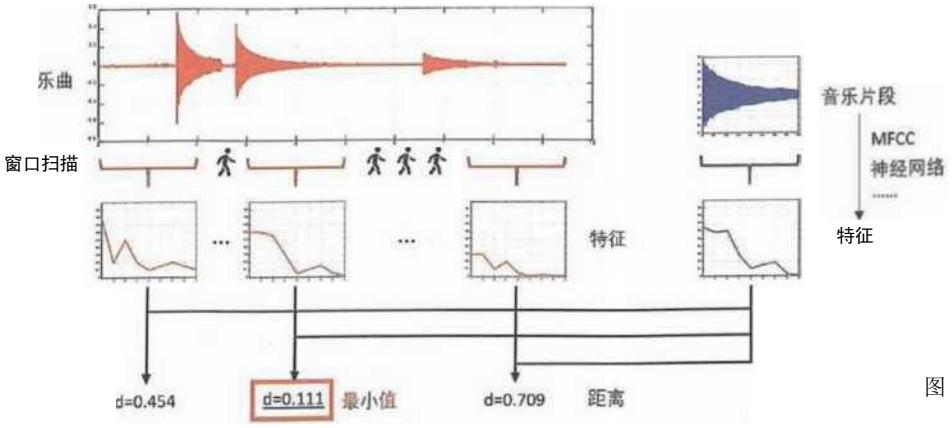


图 4-13 窗口扫描与距离计算

乐曲检索的任务有点像我们在编辑文档时用到的“查找”功能，我们也可以用同样的思路在所有曲目中查找输入的音乐片段，如果找到了无疑就是我们想要的曲目了。但是与文档中的精确查找不同，这里的查找是模糊的，比如可能是不同歌手演唱的同一曲目，虽然相似但不完全相同。所以我们并不能直接评判“找到”或“没找到”，而应该给出一个相似度。

我们通常用一个距离来度量相似度，距离越近，相似度越大。回顾第二章所学习的距离，给定两组特征 $x=(x_1, x_2, x_3)$，$y=(y_1, y_2, y_3)$，它们的距离是：

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

有了距离的概念，我们就可以进行相似度的查找了。如图 4-13,在乐曲上按照时间顺序依次截取和音乐片段长度一致的段落，相邻段落之间的时间间隔可大可小，通常要保证它们在时间上有较大的重叠，这一过程被称为“窗口扫描”。然后计算片段和所截段落的特征并算出它们的距离，取这些距离的最小值作为音乐片段与乐曲的距离。最终与音乐片段距离最小的乐曲即为检索的结果。

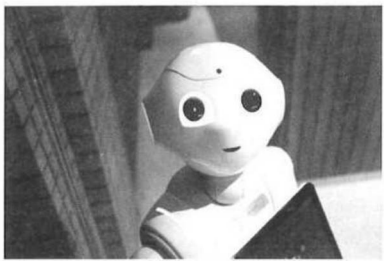
4.5 本章小结

通过本章的学习我们了解了声音的特性，并学习了如何让计算机感知与理解声音。通过声音的数字化，计算机“听”到了声音；通过频谱的计算，计算机得以理解声音的音调和音色；MFCC 特征是对频谱的再提炼，计算机可以 用更低维的向量表达出共振峰等声音的重要特性。

我们还掌握了音频理解的很多应用。借助神经网络，我们实现了对一段音乐整体风格的分类。我们还认识到声学模型和语言模型是语音识别的左膀右臂。通过最后的乐曲检索应用，我们对距离又有了更加深入的理解。

如同声音的丰富多彩，音频处理的技术也是层出不穷，这一章可以说是我们声音之旅的序曲，更多的技术和应用还在等待我们不断去探索。

第五章　冰雪聪明：看懂视频



一台名叫东东的机器人加入了铭铭的家庭，让痴迷于人工智能的铭铭兴奋异常。东东不仅聪明，还善解人意。当铭铭放学回到家中，东东立即开启服务欢迎模式，贴心地端来一杯热水；而当铭铭沉浸在学习中时，东东总是“识趣”地开启静模模式，不去打搅铭铭。铭铭一直不明白这个眨着两只大眼睛的小家伙是如何看懂自己的行为的，心里不由得赞叹其冰雪聪明。

在前面章节中，我们已经教会了计算机如何识别图像、听懂声音。在这一章，我们要教会计算机理解视频。

这些年来，互联网视频的数量日益增长，视频的内容日渐丰富，视频技术的应用也日趋广泛。面对浩如烟海的视频资源，如何让计算机自动并准确地分析其内容，从而方便我们使用呢？视频理解（video understanding）作为这一切的基础，理所当然成 为计算机视觉领域的热门方向。从光流特征到轨迹特征，从传统方法到深度学习，新方法的不断涌现推动着视频理解技术的发展。如今，无论是在视频内容分析、视频监 控，还是在人机交互、智能机器人等众多领域，视频理解技术都取得了令人振奋的 成果。

5.1 化静为动：从图像到视频

电视上播放着铭铭喜爱的体育节目，东东也在一旁安静地看着。

“东东，你能看懂吗?”铭铭调侃道。

“当然啦，电视上在播放跳水的图片。”

“怎么能是图片呢?”铭铭有些困惑，“电视上的画面明明是运动的呀。”

其实东东说得没错，人类眼中的视频本质上就是连续播放的图片。我们之所以能看到画面中的运动，是被人眼的视觉暂留机制“欺骗”了。一段跳水视频实际上是由连续拍摄的数百张照片组成的序列，其中每张照片称为这个视频的一帧（frame）。为了让同学们能够清楚地看到画面的变化，我们选取了其中具有代表性的四帧图像（如图 5-1 所示）。当几百张图像以每秒 24 帧以上的速度播放时，在视觉暂留机制的作用下，原本静止的画面就可以毫无卡顿地运动起来。化静为动，一段优美的跳水动作就这样呈现在了我们眼前。

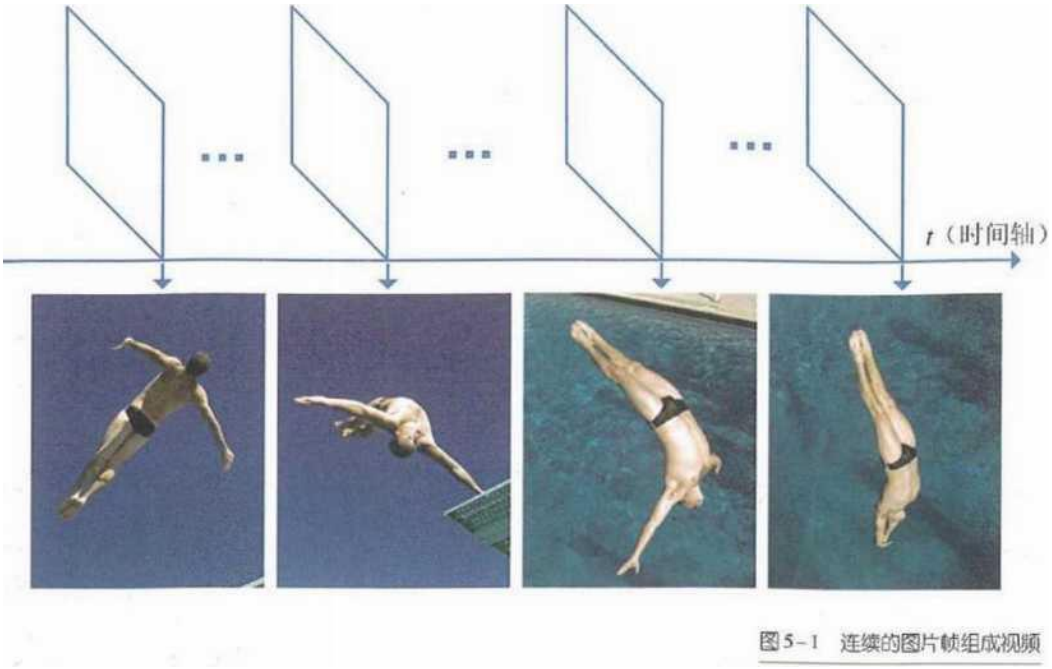


图 5-1 连续的图片帧组成视频

知识链接：视觉暂留

视觉暂留(persistence of vision)是人眼的一种机制：光在照射到视网膜后，可以保留一段时间，让人产生画面延续的印象。

我们在生活中处处都能接触到视频：电影院里放映的电影，电视上播放的节 目，DVD 存储的影片，还有手机应用中可以在线播放的视频，等等。我们在第三章中已经学习了图像在计算机中的表示方式，那么视频又是如何在计算机上表示的呢？

在计算机中，视频(video)就是按照时间顺序排列起来的图像。在播放时，只需要按照一定的速度依次将图像显示出来，就能呈现出运动的视频画面。相比于图像，我们可以认为视频多了一个维度——时间维。因此，我们可以用一个函数 I(x，y，t）来表示一个视频的信息，其中 t 是某一视频帧对应的时刻，x， y 是该视频帧中某个像素对应的位置(二维坐标)。这样的表示方法将视频和图像紧密地联系了起来，使我们可以运用图像领域的很多技术来进行视频方面的研究。了解了视频的表示，下面我们开面我们开始学习如何对视频中的行为进行识别。

5.2 明察秋毫：视频行为识别

行为（action)是人类在执行某一任务的时候所发生的一连串动作，视频行为识别（action recognition)是计算机分析给定视频数据，辨别出用户行为的过程。行为由一连串的动作组成，摄像头把动作按事情发展顺序记录下来，作为行为识别任务的输 入：行为识别的输出是给定行为集合中某个行为的名称。视频行为识别和图像分类一 样，是计算机视觉领域的基础问题。

图 5-2 视频行为识别

视频行为识别在很多领域具有重要应用价值。例如，在人机交互领域，行为识别可以让人机交互系统更精确地理解人的行为，从而给出精准的反应；在视频监控领域，行为识别可以识别监控视频中的特殊与异常行为，大大减轻警察的工作量；在基于内 容的视频索引方面，行为识别可以根据视频里的人物发生了什么行为自动把视频归类。

行为识别的挑战

人类的行为本身就是一个非常复杂的过程，让计算机理解起来有着很大的难度。不仅如此，拍摄视频时距离、光照、角度以及遮挡等因素也会给视频行为识别造成很大影响。

综合来看，视频行为识别的难点主要有以下几个方面。

首先，行为的类内差异大。类内差异，是指同一类别的行为之间存在着较大的差异。如图 5-3 所示，不同人做出的“刮胡子”行为不尽相同。如何让计算机从如此多种多样的行为表现中提取出共同的属性，是一项很有挑战的工作。

其次，行为定义的不明确导致视频缺乏代表性。如图 5-4 展示的是数据集中吃饭的视频，但是其中混杂着喂饭这一行为，这会导致计算机对吃饭行为的理解存在偏差。

最后，环境背景等差异大。如图 5-5 所示，同样是看电视的行为，从不同角度拍摄的视频环境背景有着巨大差异：一些视频中有电视屏幕的出现，另一些视频中则没有。

图 5-3 类内差异大

图 5-4 行为定义不明确

图 5-5 环境背景差异大

此外，现今行为数据集的样本数十分有限。比如在一个常用的行为数据库 UCF101 中只包含了从 YouTube 视频网站收集的 13320 个行为视频，涵盖了 101 个类别；相较之下图片数据集 ImageNet 有 1400 多万幅图片，涵盖 2 万多个类别。当然，有挑战才有发展的动力，在克服这些困难的过程中，视频行为识别技术也在逐步提高。

行为识别的重要特征：运动

经过前几章的学习，大家知道，特征的选择会对分类的准确率产生很大的影响。在进行鸢尾花的分类时，我们提取了花瓣的长宽作为区分不同品种鸢尾花的重要特 征,而如果选择花瓣的颜色作为特征，可以想象，分类器很难根据这个特征来区分不同品种的鸢尾花。既然特征的选择十分重要，那么我们针对视频行为识别该如何设计较好的特征呢？

首先我们思考一下，人类在生活中是根据什么信息来判断一个人的行为的呢？如图 5-6 所示，体育课上有的同学在练习跳高，有的同学在练习跳远，我们一眼就可以判 断出行为的类别。这是因为跳高和跳远的同学有着不同的运动过程：一个是向上高高跃起，提膝抬腿，跨过横杆后，双腿下垂落地；而另一个是向前远远跳出，双腿屈膝前探落地。可以看出，运动(motion)是我们判断行为类别的重要特征。

图 5-6 跳高、跳远动作示意图

运动的刻画：光流

区分不同行为动作的重要依据是运动，那么应该如何提取视频中的运动信息呢？ 对于我们人类来说，识别出目标在三维空间中的运动很简单。但是在计算机看来，视频只是一帧帧图片的序列。它并不知道这些图像中的人在哪，更无从知道这些目标做出了怎样的运动。这就需要我们设计一种算法，让计算机能够从序列化的图像中得到

人体的运动特征。

从物理知识中我们知道，在现实世界的三维空间中，可以用位移、速度等物理量来描述空间中一个点从一个位置经过一段时间到达另一个位置的运动过程。在视频处理中，我们用光流(optical flow)来描述运动的情况。确切地说，光流描述的是三维 的运动点投影(project)到二维图像之后相应的投影点的运动。由于我们处理的是运 动被拍摄后的二维图像数据，所以只能用二维投影点的运动来间接地刻画真实世界中的三维运动。

为了理解光流是什么，我们不妨以二维平面中运动的点投影到一维直线上的情况作为例子。如图5-7所示，t时刻一个点位于二维平面中的P_t处，经过相机的拍摄，会在一维直线l上得到它的投影点P'_t。经过时间Δt，它运动到了$P_{t+\Delta t}$点，在直线l上的投影点就移动到了$P'_{t+\Delta t}$点。向量$\overrightarrow{P'_t P'_{t+\Delta t}}$描述了点的投影点在直线$l$上的运动过程，也就近似地描述了点在真实的二维平面中的运动状态。当时间间隔Δt足够小时，向量$\overrightarrow{P'_t P'_{t+\Delta t}}$就可以看成投影点的瞬时位移（displacement），也就是我们所说的光流。

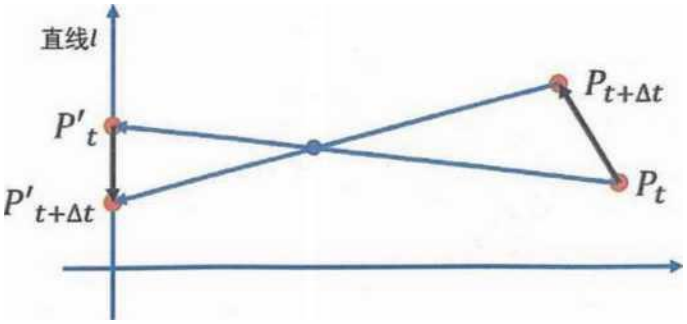


图 5-7 二维平面中运动的点在一维直线上的投影

我们看看视频中的光流怎么计算。图5-8是视频的相邻两帧$I(x, y, t)$和$I(x, y, t+1)$，由于光流就是同一个点在相邻两帧的位移，所以计算光流的关键是把两帧之间的相同的点对应起来。为了能够找出相互对应的点，我们需要两个关键的假设：(1) 相邻两帧的物体运动比较小；(2) 相邻两帧的颜色基本不变。有了这两个假设，我们就知道图像中的像素点从t时刻运动到$t+1$时刻后它的位置、颜色和亮度变化不会很大。也就是说，对于第t帧$I(x, y, t)$中的像素点$P=(x_1, y_1)$，我们只需要在第$t+1$帧$I(x, y, t+1)$中的对应位置周围寻找和像素P颜色一致的像素点$P'=(x_2, y_2)$，将P'看作P运动后到达的位置。这样得到对应的点后，就可以计算出第t帧中点P处的光流ω：$(u, v)=(x_2, y_2)-(x_1, y_1)$。

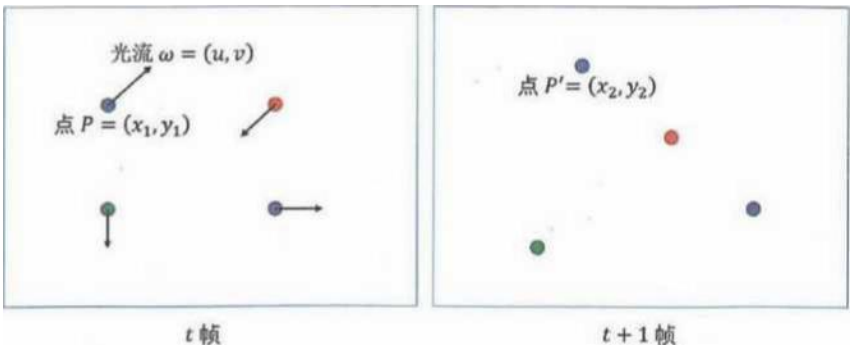


图 5-8 相邻根据相邻两帧计算光流

在实际应用中，光流的估计还需要考虑很多其他因素，比如遮挡、光照变化以及由运动产生的模糊等。因此，光流的具体实现在技术上还是比较复杂的。如图 5-9 所示，(a) (b)是相邻两帧图像。我们选取蓝色矩形框内的部分计算光流，并将每 个点的光流向量用箭头表示在图(c)中。其中，箭头的方向就是该像素点的运动方向，箭头的大小就是像素点运动的位移大小。从这个光流图可以看出，它描述的是物体向右上方的运动。

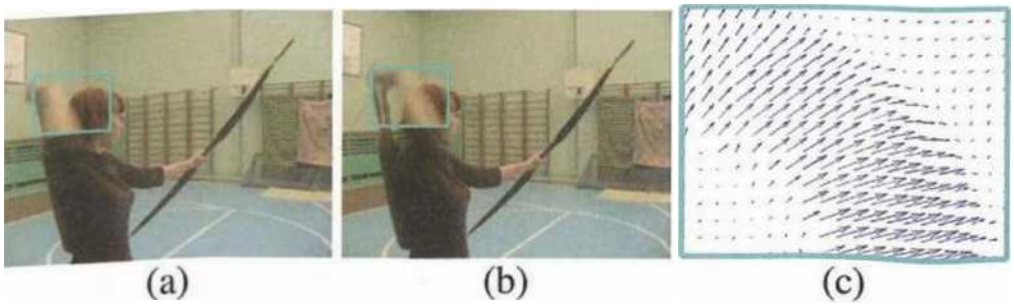


图 5-9 （c）是由相邻两帧图像(a） 和（b）计算得到的光流图

实验 5-1

在这个实验中，我们将对 UCF 101 数据集里的视频提取光流，通过观察原视频以及提取的光流图，加深对光流知识的理解。

实验步骤：

1. 观察 UCF 101 数据集，对视频的内容有一个基本的认识。
2. 使用我们提供的函数，针对视频提取光流。
3. 利用我们提供的函数，对光流进行可视化。
4. 观察光流图的特点，结合原视频，加深对光流的理解。

光流直方图

在第三章，我们学习过方向梯度直方图的概念。通过对图像中梯度信息的统计, 梯度直方图能够表示出图像中物体的轮廓信息，从而便于计算机对图像中的物体进行区分。类似地，研究者们提出了光流直方图(Histograms of Optical Flow, HOF)特征。光流直方图对视频中的光流信息进行统计，从而表示出视频中物体的运动信息，以便计算机对视频中的行为进行区分。现在就让我们揭开光流直方图特征的庐山真面目吧。

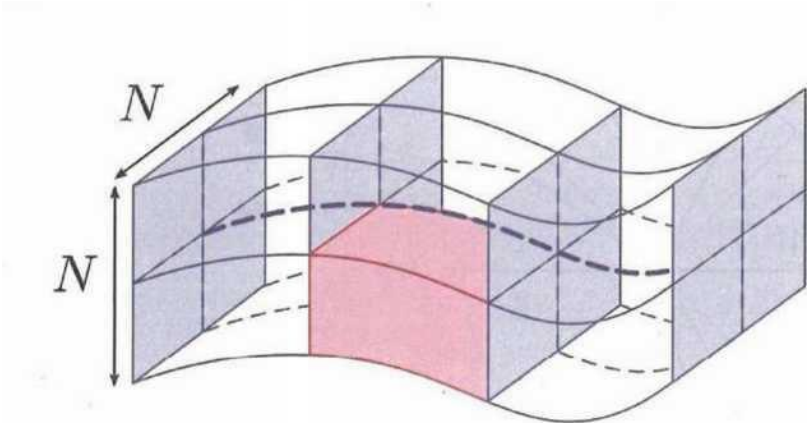


图 5-10 12 个时空单元

首先，我们把起始时刻记为 t，在 t 时刻的视频帧中选取一个点，将其位置记为 P_i=（x, y)。在 t 时刻到 t+ L 时刻的每一张视频帧中. 我们均以点（x, y)为中心截 取大小为 NxN 的区域。这样就得到了由 *L* 张同样大小的局部区域图像组成的时空体 *(space-time volume)。*

然后，我们对这个时空体进一步分割：在每一张图像上用 2x2 的网格将其划分为 4 个更小的区域，在时间维度上，我们把它分割成 3 个相等的部分。于是，我们就可以得到如图 5- 10 所示的 12(-2 x 2 x 3)个时空单元(space-temporal cell)。

接着，在每个单元内部，我们对每个像素位置处的光流进行统计。假定图像中某一像素点（x，y)的光流为 ω(x，y)=(μ，v)。这是一个二维的向量，其中 μ，v 分别表示 x 轴和 y 轴方向的光流分量。由此可得像素点（x，y)处的光流大 小为：

$$H(x, y) = \sqrt{u^2 + v^2}$$

像素点（x，y)处的光流方向为：

$$\theta(x, y) = \tan^{-1}\left(\frac{v}{u}\right)$$

为了便于统计，我们把二维坐标系中的[0°, 360。）范围划分为 8 个相等扇区， 每个扇区涵盖的角度为 45°。如图 5-11 所示，将一个时空单元内所有像素点处的光 流向量（μ，v)根据大小和方向画在上述坐标系中。然后，根据每个扇区内所包含的 光流向量进行直方图统计。比如，标号为 1 的扇形区域内包含了一个大小为 0. 8 的光 流向量，那么就在光流直方图中第 1 个位置加上 0.8；而标号为 6 的扇形区域内包含了两个光流向量，大小分别为 1. 1 和 0.6，那么就在光流頁方图中第 6 个位置加上 1.7 （即 1.1+0.6)。在统计完每个扇形区域内的光流向量后，我们就得到了该时空单元对应的光流直方图了。把这个直方图的信息用一个 8 维向量来表示，于是，我们就得到了该时空单元的一个 8 维特征向量。

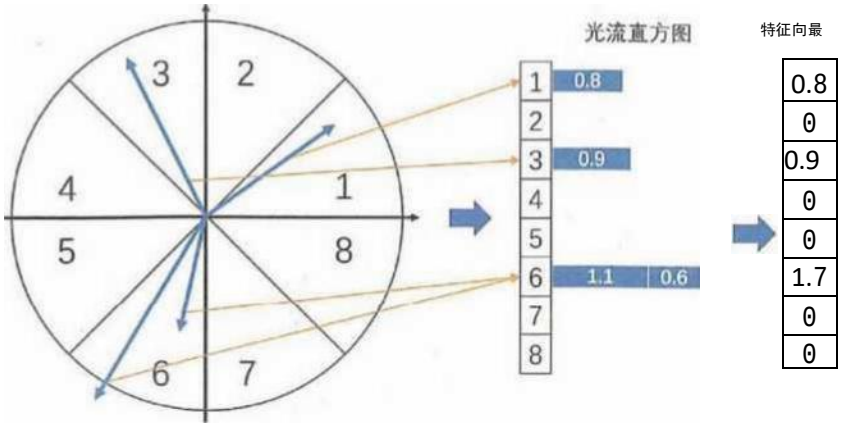


图 5-11 光流直方图

我们前面说过，一个时空体中包含了 12 个时空单元，对于每一个单元，我们都可以按照这种方式计算出一个 8 维的光流直方图特征向量。然后，我们把这 12 个 8 维向量按一定顺序拼接成一个 96 维（12 x 8)的向量，以此作为整个时空体所对应的光流直方图特征向量。

•思考与讨论•

上述例子较为简单。实际上，一个时空单元内的光流向量个数远远不止 4 个。不过，即使有更多的光流向量，我们也依然按照这样的方式进行计算。最终得到的时空体的持征向量仍旧是 96 维，同学们可以思考一下，这是为什么呢？

实验 5-2

在这个实验中，我们将利用光流直方图和支持向量机，在 UCF101 数据集上完成行为识别的任务。

实验步骤：

1. 观察 UCFIO1 数据集，对不同行为类别的视频有一个基本的认识，并区分开训练集与测试集。
2. 使用我们提供的函数，针对全部视频提取光流。
3. 利用在训练集上提取的光流直方图完成 10 类的支持向量机的训练，记录训练集上的分类正确率。
4. 利用训练好的支持向量机对测试集的光流直方图进行分类，记录测试巢上的分类正确率。

拓展阅读：从光流到密集轨迹

一个行为的发生，从开始到结束往往会延续较长的时间。我们注意到光流只描述了相邻两帧之间目标的运动，而光流直方图对时间的分割又比较粗糙。这些方式在描述运动的时间维度信息时有比较大的局限性。要更准确地刻画长时间的运动，我们就要结合连续多帧的信息。在这一小节中，我们学习一种可以描述一段时间内物体运动状态的特征--轨迹特征。

如图 5-12 所示，设第 t 帧图像中某个点的坐标为 P= (x, y), 运用光流信息， 我们就可以逐步计算出这个点在下一帧中的位置 P_{t+1}。我们知道，用一系列位移量可 以描述一个动作过程。所以，我们可以运用公式 $\Delta P_t = P_{t+1} - P_t = (x_{t+1} - x_t, \ y_{t+1} - y_t)$ 依次计算出特征点 P 在 L 帧中每一次的位移量，进而得到向量 $(\Delta P_t, \ldots, \Delta P_{t+L-1})$。我们把这个长度为 2 x L 维的向量称为轨迹(trajectory), 用它来描述一段 时间(即 L 帧)内特征点 P 的运动过程。

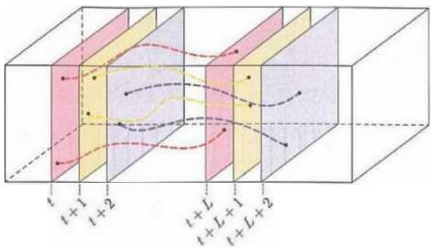


图 5-12 采样点在连续 L 帧内的运动轨迹

5.3 基于深度学习的视频行为识别

由第三章的介绍可知，卷积神经网络可以很好地提取图片的特征进行图片识别，而视频是由连续的图片帧组成，那么怎样把卷积神经网络应用到视频行为识别呢？下面将介绍近几年来主流的视频行为识别方法。

基于单帧的识别方法

当我们不考虑视频中图片信息在时间维度上的变化时，就可以用视频中的某帧图片代表整个视频的信息。

图 5-13 是用基于单帧的方法进行视频行为识别的示意图。图中最下面一层表示视频中的图片帧序列，在这里，我们随机抽取视频的某一帧图片代表整个视频，送入卷积神经网络中进行识别。当待识别的视频描述的是相对静止的行为时（例如看 电视或写作业等），只用单帧图像的特征做行为分类就能取得不错的效果。在这种情况下，各个图片帧之间差异不显著，一张图就可以代表视频的大部分信息。然而，在运动性比较强的情形下，识别行为就需要结合一连串的动作。比如跳高和跳远，在助跑阶段很相似，只用单帧的图像就很难做出区分，这会导致分类的正确率下降很多。

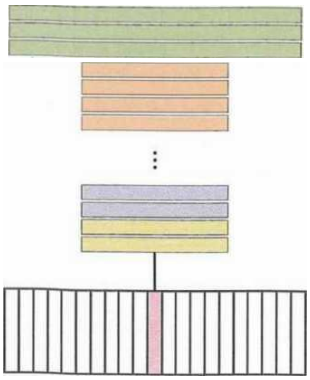


图 5-13 单帧行为识别示意图

双流卷积神经网络

由上节内容可知，当我们忽视时间维度图片帧的变化，把视频看成静止图片时，获取的是视频的静态信息，此时丢失的动态信息会影响行为识别准确率。在视频行为识别中，获取动态信息，即学习视频时间维上的表达，成为提高识别准确率的因素。这在运动性强的视频上显得更为重要。那么应该如何刻画视频中的运动信息呢？我们也许会想到利用前面介绍的光流信息，那么应该如何提取光流中的运动特征用于行为识别呢？

在前面的图 5-9 (c)中，我们看到光流在每个像素有两个分量，分别代表水平和垂直方向的位移。如图 5-14 (b)所示，如果我们把所有的水平位移取出来，然后再把它们的值缩放到 0 到 255 之间，那么就得到了一张灰度图像，叫作水平方向上的光流灰度图。同理可以得到垂直方向上的光流灰度图。水平和垂直光流灰度图，可 以作为卷积神经网络的输入，提取视频中的运动特征。

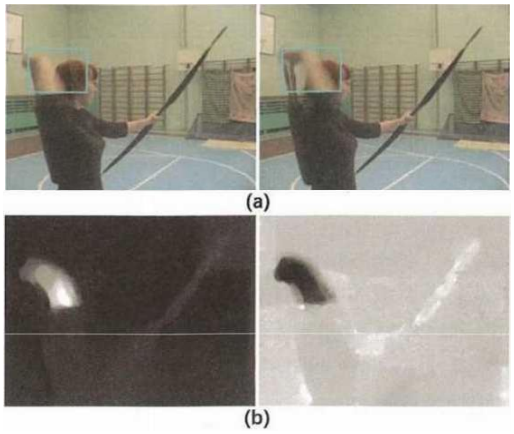


图 5-14 相邻视频帧及其光流 图(a)视频中连续的两图片帧 (b) 水平光流图和垂直光流图

此时，我们可以将视频的信息分成静态和动态两个方面，静态信息指图像中物体的外观，包含相关场景和物体，这可以通过静态图片帧获得。动态信息指视频序列中 物体的运动信息，包含观察者和物体的运动，可以通过光流灰度图来获得。视频行为识别中广泛应用的双流卷积神经网络（two-stream CNN)就利用两个不同的网络来实 现同时处理静态和动态信息。如图 5-15，我们把以随机抽取的单个彩色图像帧作为输入的网络称为空间流卷积神经网络(spatial stream CNN)而把多帧（比如 10 帧）

的光流图像作为输入的网络称为时间流卷积神经网络(temporal stream CNN)。由于双 流卷积神经网络用的是两个独立的卷积神经网络，所以在我们得到两个流对各行为的 得分后，需要采用按类别取平均或者取最大的方法，对两个流的行为得分进行融合。

我们知道，在双流识别框架中，时间流卷积神经网络是从输入的光流图中提取特 征。联想一下前面讲过的光流直方图特征，两者都是从光流中提取特征，那么它们在 方法上有什么区别吗？事实上，光流直方图是对光流方向进行加权统计，得到光流方 向信息直方图，属于手工设计的特征。而在双流识别框架中，是将光流分解为水平和 垂直光流图，然后送入卷积神经网络提取出运动特征。这是让计算机自动地学习出光 流图中的运动信息。与手工设计的光流直方图相比，卷积神经网络具有更高效的特征 表达能力，通过网络逐级提取从底层像素到高层语义的特征，从而更有效地进行行为 识别。

r识别示意图

值得注意的是，这里光流图的堆叠是为了捕捉时序相邻帧之间的运动信息，如果 输入的帧数太少，会造成时序信息捕捉不完全，无法代表较长的视频序列。如果输入 的帧数太多，又会造成计算量的增加。那么输入多少帧合适呢？实践经验表明，当光 流图帧数选取到一定值后，帧数对识别的精度影响不大，所以可以选用此临界值，它 平衡了对精度和速度的要求。

实验 5-3

在这个实验中，我们将用双流卷积神经网络代替传统特征和模式分类系统， 同样在 UCF 101 数据-集上完成视频行为分类任务。时间流和空间流神经网络都选 用 VCG16 网络。

实验步骤：

1. 利用工具包中提供的函数对 UCF 101 中训练集与测试集的视频分别进行提帧和提光流操作。

2. 利用工具包中提供的函数和从 UCF 101 训练集中提取的图片帧与光流图分别进行时间流和空间流网络的训练，分别记录两个网络在训练集上的分类正确率并比较。

3. 在 UCF 101 的训练集上，利用工具包中提供的函数，探究不同帧数的连续光流图对时何流神经网络识别准确率的影响。

4. 如果将时间流网络和空间流网络的输出按照 l ：1 加权结合，不一定能得到最高的测试准确率。对不同的实验设定，最优的加权结合比例也不同，我们寻找当前情况下的最佳融合比例。

5. 利用 UCF 101 的测试集对训练好的双流卷积神经网络进行测试，并记录其在测试集上的分类正确率。

6. 比较双流卷积神经网络和传统模式分类系统的分类正确率。

长视频的处理：时序分段网络

对于短视频（10 秒左右），双流神经网络可以很好地进行识别。对于长视频的识别，它给我们带来的挑战是如何对较长的时间进行建模。这里我们介绍另一种处理长视频数据（几分钟）的神经网络时序分段网络（temporal segment networks, TSN)），它是为了解决视频中存在跨越时间长的行为被提出的。

为了获取视频中长时间内的运动信息，如果对时间流进行密集（dense)）的采样，则需要更多的光流帧来覆盖运动的始末。过多的光流帧带来的问题是产生过多的计算量，不适合进行应用推广。此外，由于视频中的连续帧冗余较多，如图 5-16 所示，相邻帧之间极为相似，若进行密集采样，则采样的图片之间会极为相似。若在相邻帧间进行时间上的稀疏（sparse)采样策略，不仅会节省计算成本，而且也不会遗漏视频的重要信息，相比于密集采样更加可行。

图 5-16 连续的视频帧

稀疏时间采样(sparse temporal sampling)策略，就是对于长度不同的数据，根据时间先后分成固定数量的段落。好比不论每个班有多少个同学，排座位时都分成 6 组，如果同学数量比较多，那么每个组的人数就相应增加，反之则减少。从每个段落提取一个特征，可以得到固定长度的特征，后面再接一个处理固定长度数据的网络就 可以了。从一个视频段落提取一个特征有很多种方式，从前面的介绍可知，可以是最简单地抽取一个样本帧，或者用双流卷积神经网络。时序分段网络就好像让班里每组同学提交一份作品，可以选一个同学完成，也可以对每个同学进行不同的分工，大家合力完成。

图 5-17 长视频实例

对于做饭这一复杂的行为，会有一系列连续的动作组成，也存在镜头转换，属于跨越时间较长行为的视频，最终的行为得分应该结合各个时间段的行为得到整个视频 的行为类别。

图 5-18 是时序分段网络的示意图，具体做法是，在时间上将一个输入视频划分为几段(图中分成了三段)，并从每个片段中随机选择连续光流图和图片帧，对每个片段都用双流卷积神经网络的框架进行行为识别；最后将三个片段的类别得分进行融合，得到整个视频的行为类别。每个片段都对视频的预测产生贡献，每个片段先产生自己的初步预测动作类，然后结合各个片段的预测，融合后得出对视频整体上的预测。

时序分段网络解决长视频识别的关键在于把视频沿时间轴分段，使得采样样本能较为均匀地分布于整个时间段。因此网络能够模拟整个较长时间的结构，模型也能够动态覆盖整个视频。

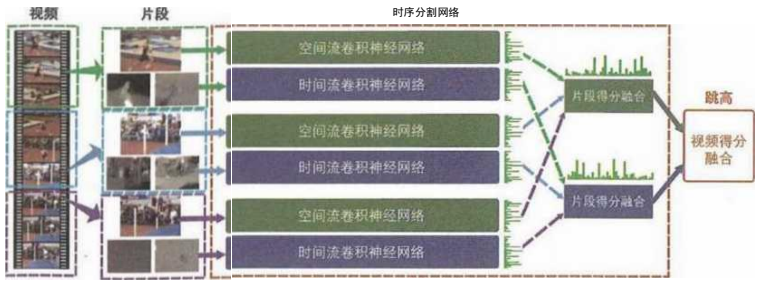


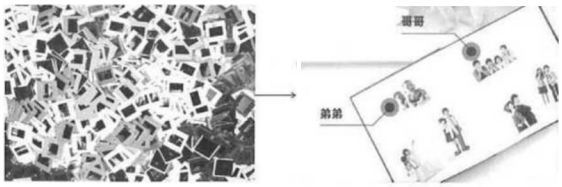
图 5-18 时序分段网络示意图

5.4 本章小结

通过这一章对视频行为识别的介绍，我们理解了视频和图片的区别与联系，认识到了对视频的时间结构进行建模的重要性，还掌握了如何对视频的时空特征进行有效的特征提取与表示。具体来说，光流直方图和轨迹分别以不同方式刻画了视频中行为的运动特征，从而为视频行为识别奠定了基础。在深度学习方法中，我们利用卷积神经网络对视频的光流特征等进一步进行有效的提取和表示，并用双流框架结合视频运动信息和外观信息进行分类。最后，针对长视频中的行为识别任务，我们简单介绍了稀疏时间采样策略以及时序分段网络。

视频综合了图像、音频、文本等多种形式的信息，如何进一步结合各个方面对视频进行分析和理解还有待同学们去探索和研究。

第六章 无师自通：分门别类



照片　　整理好的相册

从铭铭出生的那一天起，爸爸就开始用数码相机记录下铭铭的点滴生活长大后的铭铭喜欢 打开爸爸的老相机，看看年轻时的爸爸妈妈，调皮的自己，还有不再联系的小伙伴。只是，相机里除了铭铭认识的人，还掺杂着爸爸的同事、妈妈的朋友，以及随手拍下的风景照。铭铭一边按着翻页键跳过这些无趣的相片，一边想：“计算机能不能自动地将照片整理好呢？”

6.1 当人工智能未曾听说花的名字

图 6-1 加斯帕半岛的鸢尾花

大约在 100 年前，加拿大加斯帕半岛的 150 株鸢尾花得到了植物学家埃德加•安德森的测量。安德森先生仔细地记录了这 150 个幸运儿的外表特征与品种。这份数据流传至今，成为许多人工智能第一次认识世界的时候使用的“启蒙读物”。在第二章里，分类器也正是从这份数据中学会了区分鸢尾花的品种。

可以认为，分类器对鸢尾花的认识，来自安德森的植物学知识。因为分类器是在安德森提供的品种标注信息的指导下，才学会区分鸢尾花品种的。这种需要训练数据的标注信息的学习过程，就是监督学习。我们回忆，前面章节中对鸢尾花、图像、音频和视频的分类，也都需要类别的标注信息。它们都属于监督学习。

尽管人类历史可以追溯到数百万年前，但直到三百年前，我们才拥有一套科学完备的生物分类体系。这套体系的提出者，正是现代生物分类学之父卡尔•冯•林奈 (Carl von Lime)。在那个大量生物未被命名的时代，年轻的林奈在他的笔记里写道：“生物与生物之间，生物与大自然之间，为何充满着这么有趣的关系?”倘若人工智能独自到达加斯帕半岛，它能否像充满好奇心的林奈一样，探索生物间的关系；倘若人工智能未曾听说鸢尾花的名字，它能否独立地为鸢尾花分门别类呢？

与第二章的鸢尾花分类任务有所不同，为生物划分门类时，并没有类别的标注信

息供人工智能参考。

我们称这种没有标注信息的学习过程为无监督学习。图 6-2 左图是监督学习中鸢尾花数据在人工智能眼里的样子。在类别信息的指导下，我们很容易就能找到一条最优的直线，将特征空间一分为二，使变色鸢尾和山鸢尾各居一隅。而图 6-2 右图则是无监督学习的情况。可以看到，没有类别信息的指导，我们很难判断哪一些鸢尾花是相同品种，哪一些鸢尾花是不同品种，更别提使用一条直线为鸢尾花分类了。

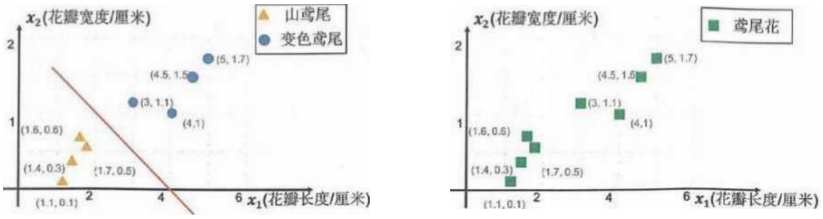


图 6-2 人工图能眼中的鸢尾花：左图与右图分别表示监督学习与无监督学习的情况

幸运的是，尽管没有指导信息，但我们知道. 同一品种的鸢尾花，它们的花瓣宽度与花瓣长度应该是相近的。也就是说，特征空间里相近的两个样本点，很可能是同一种鸢尾花。从图 6-3 左图可以看到，特征空间里的鸢尾花大概聚集成了两簇。属于同一簇的两朵鸢尾花，比如 α 与 β，拥有相似的花瓣宽度与花瓣长度。而属于不同簇的两朵鸢尾花，比如 α 与 γ，花瓣的长度相差甚远。从直觉上说，这两簇鸢尾花很可能正好属于不同的两个品种。如图 6-3 右图所示，我们分别记这两簇鸢尾花为 A 类鸢尾花与 B 类鸢尾花。

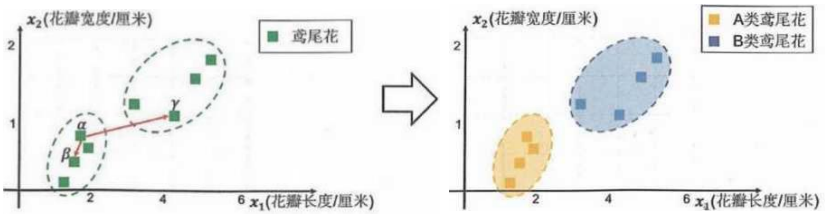


图 6-3 根据鸢尾花花在特征空间的聚集情况，也可以进行分类

可见，通过分析数据在特征空间的聚集情况，也可以将一组数据分成不同的类。 我们称这类方法为聚类(clustering)。聚类旨在把一群样本分为多个集合，使得同一 集合内的元素尽量“相似”或“相近”。聚类的一个重要假设，就是特征空间里相近的两个样本，很可能属于同一个类别。这一假设不一定在所有数据中都成立，我们在使用聚类算法的时候，应该特别注意这一点。聚类作为一种无监督学习过程，不需要数据的类别标注，甚至不需要预先定义类别，是一种非常实用的分析方法。

在这一章里人，工智能将像真正的科学家一样，自己从鸢尾花数据中发掘规律， 发现不同品种的鸢尾花。

6.2 物以类聚：鸢尽花的 K 均值聚类

【问题】有 N 株鸢尾花，第 n 株鸢尾花在特征空间的坐标为

$$x_n = (a_n, b_n),\ n = 1, 2, 3, \ldots, N$$

其中，a_n, b_n 分别代表第 n 株鸢尾花的花瓣宽度、花瓣长度。我们希望人工智能在不知道鸢尾花品种的前提下，将这 N 朵鸢尾花分为 K 类，使得同一类样本的特征相似程度高，而不同类样本的特征相似程度低。下面我们介绍一种聚类方法。它的主要思路是先从任意一组划分出发，通过调整，逐步达成上述的目标。

对于如图 6-4 所示的划分方式，计算每一类鸢尾花花瓣的平均宽度与平均长度， 我们可以得到每一类鸢尾花的聚类中心(cluster center),它们位于图中的红色记号处。由于聚类中心是由一类鸢尾花的平均特征决定的，它可以作为该类鸢尾花的代表。一株鸢尾花到某一类鸢尾花中心点的距离越小，就表示它与该类鸢尾花越相似，越可能属于该类鸢尾花。

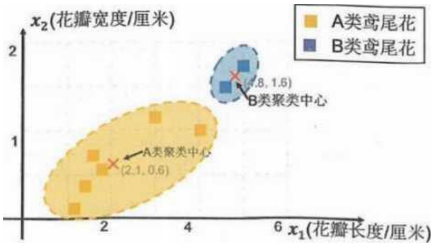


图 6-4 每一类鸢尾花的聚类中心

观察图 6-5，我们发规 A 类鸢尾花 γ 与 B 类的聚类中心更接近，这说明它与 B 类鸢尾花更为相似。这有悖于聚类的目标：同一类样本的特征相似程度高，而不同类样本的特征相似程度低。解决方法很简单，将鸢尾花 γ 归入 B 类就可以了。

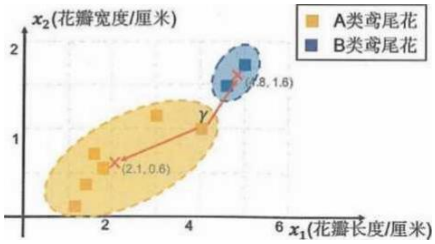


图 6-5 矛盾的样本丫

由于划分方式的变化，类别的聚类中心改变了，我们重新计算聚类中心。结果如 图 6-6 所示。然而，在新的划分方式中，我们又发现了新的矛盾：A 类鸢尾花 ω 更 接近 B 类鸢尾花的聚类中心。为了进一步改善划分结果，我们将鸢尾花 ω 归入 B 类。

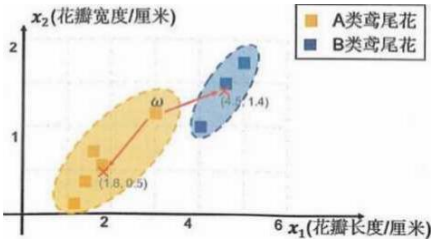


图 6-6 矛盾的样本 3

如图 6-7 所示，经过多次修正，我们终于得到一个令人满意的划分方式，所有的鸢尾花都满足聚类目标。

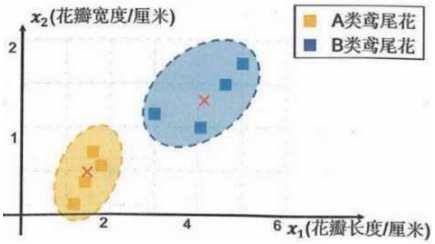


图 6-7 K 均值聚类结果

上述的聚类方法，就是 K 均值聚类（K-means clustering)算法。在 K 均值聚类算法中，已知样本的划分方式，可以计算每一类样本的聚类中心。反之，已知聚类中心，也可以得到一个更好的划分方式。通过循环地改善中心点与 划分方式，我们可以得到越来越好的聚类结果，直到聚类中心与划分方式不再发生变化。

那么，最初的聚类中心要如何获得呢？首先，我们要决定聚类的类别数 K，再在所有样本中随机选取 K 个样本作为聚类中心，就完成聚类中心的初始化了。完整的 K 均值聚类算法如下：

K 均值聚类算法：

第一步，随机地从所有样本中选取 K 个样本，作为每一个类别的初始聚类中心。

第二步，将每一个样本划分给距离最近的聚类中心对应的类别，得到新的划分方式。

第三步，重新计算每类样本的聚类中心。

重复第二、三步骤，直到聚类中心与划分方式不再发生变化。

下面仍用鸢尾花数据集来演示 K 均值聚类算法。在第二章以及前面的例子里，我们都只用了这个数据集中的山鸢尾和变色鸢尾两类数据。其实完整的数据集中还包括弗吉尼亚鸢尾 ***(Virginia iris)***，共三类。为了增加问题的难度，我们用全部的二类数据（隐藏类别信息不用）进行聚类。图 6-8 展示了 K=3 的情况下在鸢尾花数据集中 K 均值聚类的过程。圆形表示鸢尾花样本，方形表示每一类鸢尾花的聚类中心，颜色则标示了聚出的不同的类别。可以看到经过三次更新，鸢尾花被划分为相对集中的三簇。

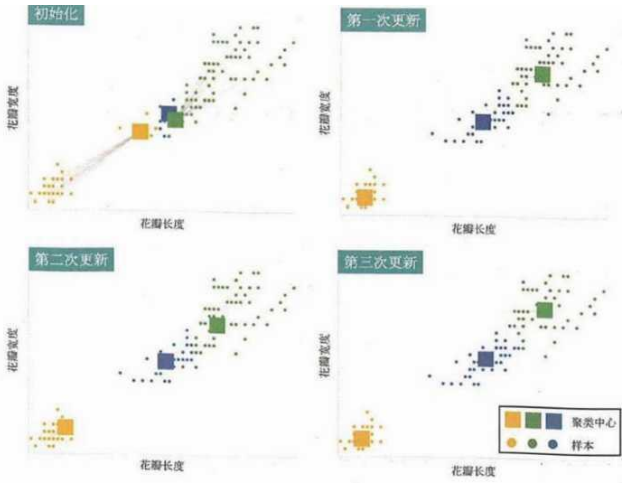


图 6-8 K 均值聚类的收敛过程

经过 K 均值聚类后，样本被划分为 A、B、、、C 三类。每一类包含的不同品种的鸢尾花比例如图 6-9 所示：

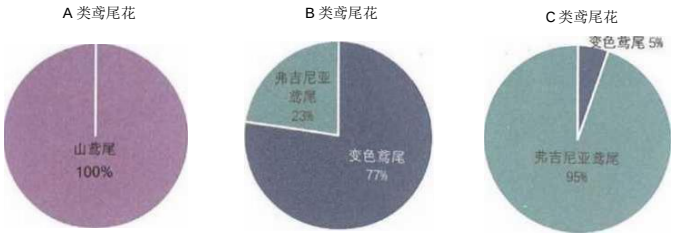


图 6-9 K 均值聚类结果的组成成分分析

其中 A 类鸢尾花全部由山鸢尾组成，B 类鸢尾花主要由变色鸢尾组成，而 C 类鸢尾花主要由弗吉尼亚鸢尾组成。 尽管我们并未提供鸢尾花的品种信息，K 均值聚类算法通过分析样本的特征，依然发现了三种鸢尾花的存在，并将鸢 尾花大致正确地划分成了三类。这是本书中人工智能第一次不依赖人类的知识，第一次独立观察世界，第一次得到 只属于自己的答案。

实验 6-1

1. 根据本节描述的步骤，以鸢尾花的花瓣长度、花瓣宽度为特征，使用 K 均值聚类将鸢尾花分为三组。统计聚类结果中的每一类各包含了多少朵山鸢鸢、变色鸢尾和弗吉尼亚鸢尾。

2. 安德森除了记录鸢尾花的花瓣长度与花瓣宽度，还记录了萼片长度与萼片宽度。同时使用这四个特征对鸢尾花进行 K= 3 的 K 均值聚类。统计聚类结果中的每一类各包含了多少朵山鸢尾、变色鸢尾和弗吉尼亚鸢尾。

3. 这两个小实验的聚类统计结果有何不同？为什么？

6.3 人以群分：相册中的人脸聚类

从牙牙学语开始，铭铭留下了许多合影。这些珍贵的照片记录着铭铭与家人、朋友的回忆。然而，当他希望重 温与某一个人的往事，或者只是想要给自己挑一张照片作为头像的时候，却不得不将所有的相片翻看一遍，整理照片费时费力。现在，人工智能将帮助铭铭整理相册，自动地将相片按照被拍摄者分组。

图 6-10 相册聚类的流程

在前文，我们将鸢尾花看成特征空间中的特征点，而后使用 K 均值算法对鸢尾 花的特征进行聚类。同样地，只要我们能对照片中的人脸提取特征，用特征空间里的 特征点表示每一张人脸，就能使用 K 均值算法将“相似”的人脸聚集起来了。

特征的提取对于聚类效果而言至关重要。要怎样获得人脸的特征呢？如图 6-10 所示，给定一个相册，对相册中的每一张照片分别进行人脸检测、人脸转正与特征提 取. 我们就获得了用于人脸聚类的人脸特征了。下面，我们将按照图 6-10 所示的流程，一步步地对图 6-11 所示的照片提取人脸特征。

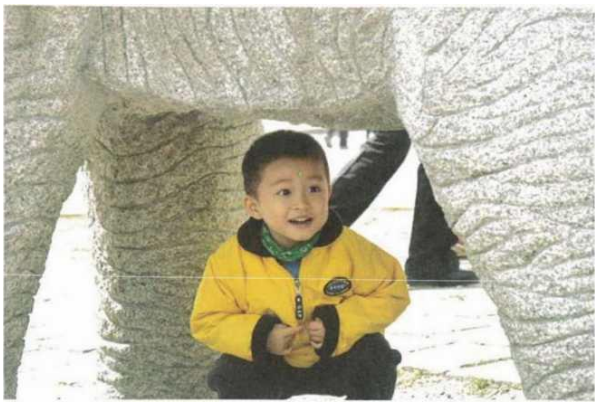


图 6-11 一张包含人的照片

人脸检测

人脸检测的目的，是定位图像中人脸的位置。在对相片进行分组的时候，我们只关注照片中出现的人物，因为“人脸”的特征正是用来辨别人的身份的重要信息。 而照片背景、人物穿着等无关信息则会干扰我们的判断。因此，如图 6-12 所示，我们首先使用预先训练好的人脸检测器找到照片中人脸的所在位置。在之后的步骤中，

仅对包含人脸的区域进行分析。

人脸检测器的原理，我们已经在第二章的拓展阅读部分讲述过了。人脸检测是一项非常成熟的技术，有许多性能出众、抗干扰能力强的人脸检测器可以直接使用。

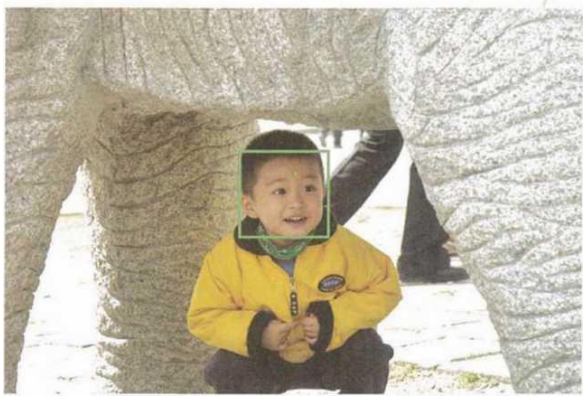


图 6-12 人脸检测：绿框表示人脸

人脸转正

人脸转正的目的，是使姿态各异的人脸统一面向正前方。在人脸聚类中，同一个 人的不同照片的特征越相似，聚类算法越容易将这些照片分为一组。然而，如图 6-13 所示，尽管人脸检测已经帮助我们将关注的区域缩小到人脸的部分，但人脸的朝向依然会使得同一个人的照片看起来大不相同，从而干扰特征提取的结果。

图 6-13 姿态不同的人脸

为了解决这个问题. 我们先找到人脸上的关键点（比如眼睛、鼻子、嘴角）。而后，我们根据关键点的位置，对图片进行合适的几何变换（如缩放、拉伸、切变）， 使脸部统一面向正前方。图 6-14 展示了人脸校准与人脸转正的过程，可以看到，图中的人脸从朝向斜前方变为朝向正前方。人脸校准在很大程度上帮助我们抵消了脸部姿态的干扰。

图 6-14 人脸校准与人脸转正：绿点表示人脸关键点

与人脸检测相似，人脸校准与人脸转正在人工智能领域也属于非常成熟的技术，可以由计算机自动完成。

特征提取

去除了脸部姿态的干扰，我们就可以使用神经网络在每个关键点附近抽取特征了。是不是任意的神经网络都可以帮助我们抽取特征呢？

在前面的章节，我们学会使用神经网络进行分类。在训练过程中，根据分类任务的不同，神经网络的每一层都在自动地寻找最适合分辨样本类别的特征。比如在第三章的图片分类任务中，神经网络寻找到的特征就包含了最能区分图片内物体的类别信息。而在人脸识别任务中，神经网络被要求判断人脸主人的身份。它寻找到的特征就更能区分人脸之间的不同。

在我们的相册聚类任务中，我们更关心照片里的人的身份，而不关心照片里出现了哪些物体。因此，我们应该选择用于人脸识别的神经网络对人脸进行特征提取。

图 6-15 展示了特征提取的过程。我们将人脸图片送入卷积神经网络，取网络中倒数第二层的输出作为描述该人脸的特征。

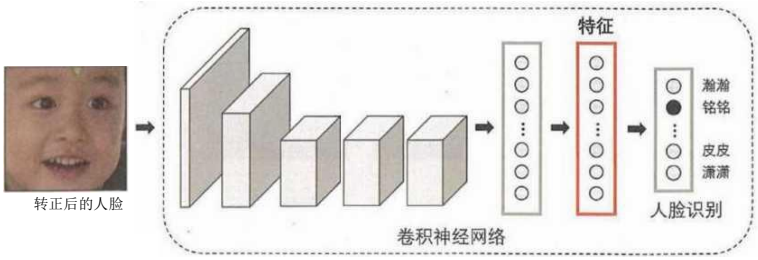


图 6-15 特征提取：红色边框所示特征是我们提取到的特征

注意，这里用于提取人脸图片特征的神经网络是在其他数据集上预先训练好的模型。这个网络在训练时虽然没有见过铭铭和他的朋友，但它提取出的特征仍然可以用来对这些新面孔进行聚类。这是因为一个好的神经网络会有比较好的推广能力。

•思考与讨论-

我们一般选取神经网络倒数第二层的输出作为特征而不用最后一层的输出。这是为什么？

实验 6-2：提取照片中的人脸特征

1．收集你与家人、朋友的电子照片，将其导入计算机。也可直接使用你的家 庭相册、班级相册或实验平台提供的相册。

2. 将照片输入提供的人脸检测器，得到人脸的位置以及关键点的位置。

3. 将照片中人脸的部分截取出来，去除背景部分，得到人脸图片。

4. 使用关键点对人脸图片进行人脸较正。将转正后的人脸图片输入到用于特 征提取的神经网络中，得到人脸特征。

人脸聚类

提取完每张照片中的人脸特征之后，我们就可以对人脸进行 K 均值聚类了。

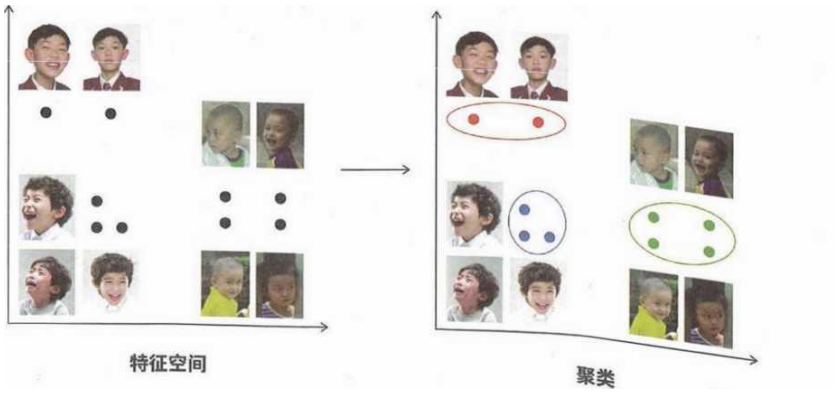


图 6-16 人脸聚类

如图 6-16 所示，与鸢尾花聚类相似，人脸聚类就是挖掘特征空间中相近的人脸的过程。通过人脸聚类，我们将相册中出现过的人脸划分为若干类。由于每一张人脸图片都是从某一张照片中截取出来的. 照片也自然地被分为若干类。注意，有些照片中包含了多个人脸，因此同一张照片可能会被划入多个类中。使用计算机对于同一类的照片进行自动排版，就可以得到如图 6-17 所示的漂亮的相片集了。

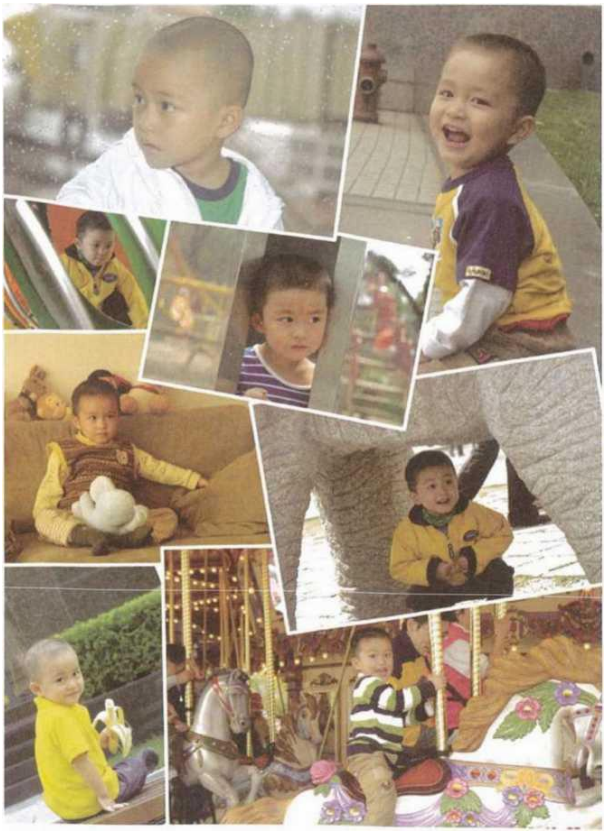


图 6-17 相册的人脸聚类的结果

在进行 K 均值聚类之前，我们首先需要确定聚类数量 K 的大小。在鸢尾花的聚类中，我们或许可以肉眼观察数据的聚拢程度，猜测鸢尾花大致应该被分为几组。然

而，在相册的人脸聚类中，我们往往不知道数据应该分为几类. 尤其是在难以直接观 测和统计的高维数据空间当中。那么，如何确定 K 的大小呢？我们使用不同的 K 进 行 K 均值聚类，统计 K 取不同值的时候每一个样本和对应聚类中心的平均距离。

样本到对应聚类中心的平均距离一定程度上可以衡量聚类的效果。从图 6-18 可 以看到，随着聚类数量 K 的增加，平均距离在不断地下降。然而，聚类数量过大，会 导致照片划分得过细，每一类中只包含很少的照片，这样就失去实用性了。如何在平 均距离和聚类数量间取得平衡呢？从图 6-18 可以看到，在 K =3 的时侯，曲线产生了一个明显的拐点。在拐点之后，随着 K 的增加，平均距离减少得非常缓慢。因此， 拐点处 K =3 是一个合适的选择。由于这条曲线与手肘的形状相似，这种方法被称为 手肘法(elbow method)，而这个拐点则被称为手肘点。

实验 6-3：相册中的 K 均值聚类

1．使用提供的 K 均值聚类代码，观察在聚类过程中每一次迭代，相册聚类结果的变化。

2．观察 K 均值聚类稳定后的结果，看看每一类都代表了什么。(注意：可能 会有一些类别代表同一个人的不同姿态，甚至是同一姿态的不同人。)

3．使用手肘法为相册聚类确定合适的 K，肉眼观察不同的 K 下的聚类结果，看看手肘法确定的 K 是不是最合适的。

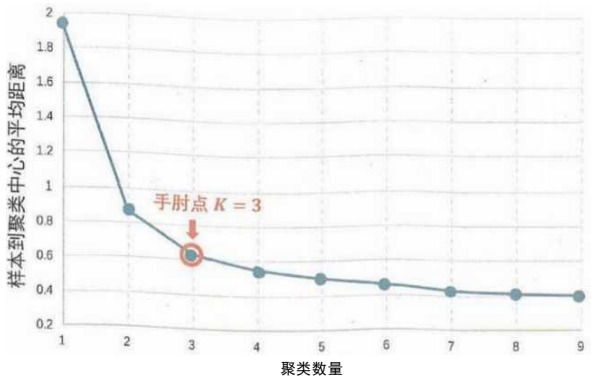


图 6-18 手肘法确定聚类数量 K 的大小

6.4 层次聚类与生物聚类

除了 K 均值聚类，层次聚类也是一种经典的聚类算法。层次聚类首先将每个样 本都单独当成一类，而后重复地合并最相似的两个类。当所有的类别间的距离都超过 一个预设的截止距离时，层次聚类就完成了。

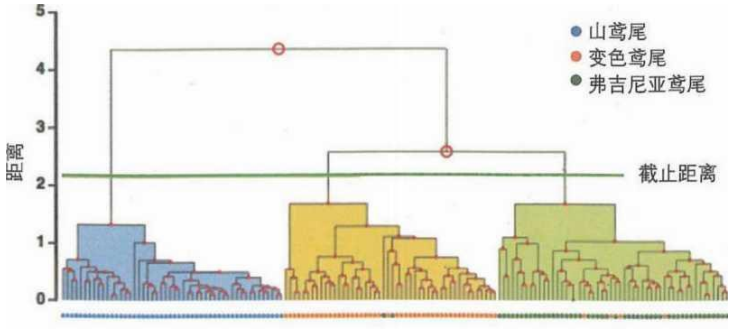


图 6-19 使用层次聚类算法对鸢尾花进行分类

图 6-19 表示了层次聚类对鸢尾花进行聚类的结果。可以看到，层次聚类得到了构筑于所有样本之上的树状图。越底层的聚类结果分类越细，越顶层的聚类结果分类越粗。而表示截止距离的绿线下方的聚类结果，正好与鸢尾花的三个品种吻合。

在生物学上. 层次聚类除了可以帮我们分辨不同品种的鸢尾花，还可以用来分析基因，推导动植物的分类甚至是进化过程。以生物 DNA 序列作为特征，不断合并基因相似度高的物种，我们就可以得到生物的“分类树”。

有趣的是，由层次聚类得到的“分类树”与图 6-20 所示的生物进化树非常吻合。随着基因测序工程的进行与无监督学习的发展，未来，人工智能还将帮助我们发现更多生物间的未知联系。

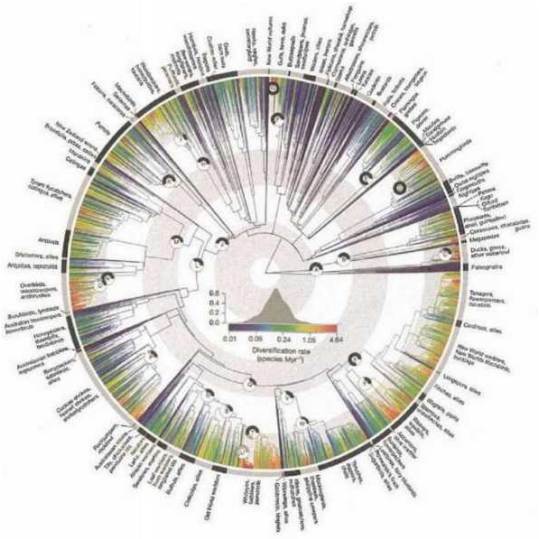


图 6-20 生物进化树

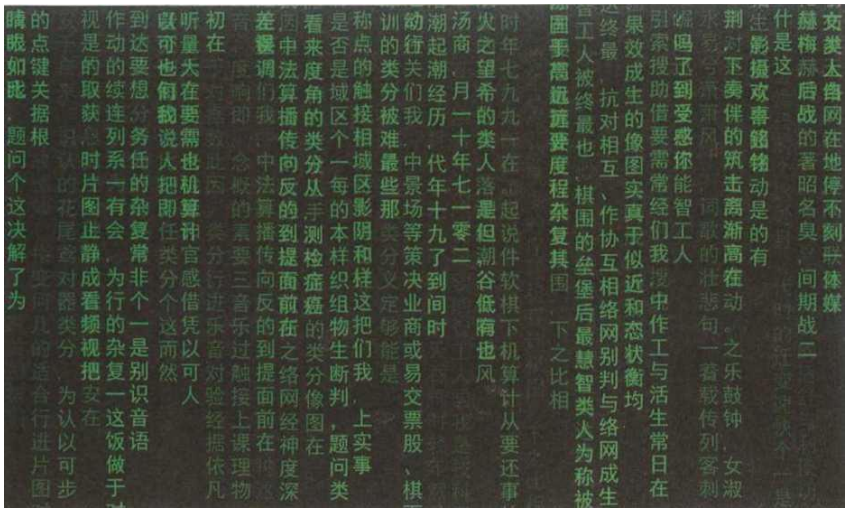
6.5 本章小结

在这一章我们学习了本书第一种无监督学习方法。与监督学习不同，无监督学习需要在没有标注信息的情况下完成对数据中的规律进行发掘。本章中学习的 K 均值 算法作为一种最基本的无监督学习算法，可以在无标注的前提下对数据进行聚类。由于 K 均值算法等聚类算法简单有效，在金融、医疗、大数据挖掘等领域都有着重要 应用。除此之外，我们在之后两章中，还将了解到主题模型、生成对抗网络等无监督学习方法。

如今互联网中充斥着大量的数据，它们以图像、文本、音频、视频等形式存在。 绝大部分数据缺乏标注信息，无监督学习也因此得到越来越广泛的重视。无监督学习的蓬勃发展，将会在未来给人工智能带来质的飞跃。

第七章 识文断字：理解文本

自人类文明诞生以来，文字就是人们传递信息的基本媒介。在互联网高度发达的今天，文字形式的信息也以爆炸式的速度增长着。媒体一刻不停地在网络上发布着最新的新闻，人们随时随地通过手机谈论着身边的事情。每时每刻都有大量的文字从各种渠道生产出来，汇集在互联网上而对海量的文本数据，我们又能否利用人工智能技术自动时其进行分析与理解，从而节省人类有闲的阅读时间与精力呢



在这一章中，我们将学习潜在语义分析(latent semantic analysis)技术。借助这项技术，计算机就可以从海量的文本数据中自动发掘出潜在的主题，进而完成对文本内容的概括与提炼。在正式介绍相关技术之前，让我们先来探讨一下“从文本中发掘 潜在主题”这个任务的特点。

7.1 任务的特点

文本数据通常不会包含额外的标注信息，例如，我们在社交网络上发布了一条消息：“我在学校学习了人工智能课程。”这句话是围绕“学习”或“人工智能”等主题展开的，但我们在发布这条消息时并不会特意将这些主题标记上去。如果我们希望 对该社交网络上的所有消息进行分析，那么能获取到的信息通常就只有消息本体，而没有如何额外的标记。

我们又能否通过人工标注的方式获得关于文本主题的信息呢？这通常也不太可能。文本数据的规模通常远大于视频、图像等多媒体信息。新浪微博 2012 年第二季 度的公开数据显示，网站每天都会产生 1.17 亿条微博。对于如此规模的数据，人工

标注的代价过于高昂。在这种情况下对数据进行分析止是九监督学习算法的用武 之地。

既然是无监督学习的任务，那么我们能否用上一章介绍的 K 均值算法，对文本数据进行聚类，从而提取出潜在的主题呢？这听起来是可行的，但我们却忽略了文本 数据的一个特点。在 K 均值算法中，我们会将一个样本划归为一个特定的类别，而一段文本通常可能围绕多个主题展开。例如，一篇关于“推动中小学人工智能教育”的新闻至少会围绕“人工智能”和“中小学教育”两个主题展开，我们将其划归为任意单一主题都是不合适的。

而潜在语义分析技术就是针对文本数据“多主题”的特点而设计的。这种技术可以通过无监督的方式从文本中分析出多个潜在的主题，完成聚类算法不能完成的任务。

为方便讨论，我们现在介绍一些相关的专有名词。我们通常将上文中提到的海量文本数据称为语料库（corpus), 语料库中独立的文本称为文档（document)，文档的中心思想或主要内容称为主题（topic)。例如，2017 年的全部报纸文章可以组成一个 语料库，报纸上的每一篇文章构成一篇文档. 这些文档可能围绕“政治”、“经济”、“ 教育”、“科技”、“民生”等主题展开。

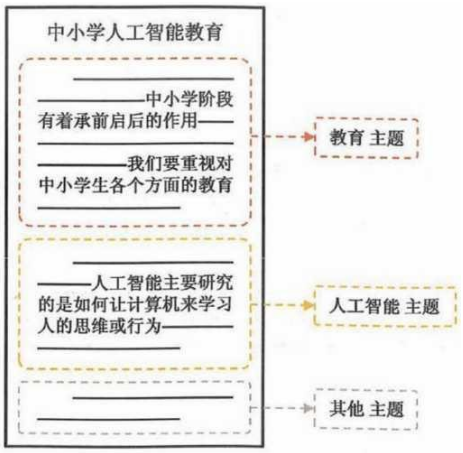


图 7-1 一篇含有多个主题的 文档

7.2 文本的特征

词袋模型

词袋模型（bag-of-words mode)是用于描述文本的一个简单的数学模型，也是常用的一种文本特征提取方式。词袋模型将一篇文档看作是一个“装有若干词语的袋子”，只考虑词语在文档中出现的次数，而忽略词语的顺序以及句子的结构。

例如，对于下面这段文本：

铭铭喜欢打篮球，也喜欢打乒乓球。

我们可以将其表示为一个由形如（词语：出现次数）的二元(tuple)组组成的 集合：

{（铭铭：1））(喜欢：2）（打：2）（篮球：1)）（也：1）)（乒乓球：1））}

这个集合就是这段文本对应的“词袋”。

词袋模型对文档进行了很大程度的简化，但一定程度上仍然保留了文档的主题信息。我们根据词袋中“铭铭”、“篮球”、“乒乓球”等词语，仍然可以知道这篇文档 与铭铭以及体育两个可能的主题相关。忽略难以建模的词句结构、保留体现主题的词语计数，便是词袋模型的基本思想。

有了词袋之后，我们可以构造一个包含若干词语的词典（vocabulary)，并借助这个词典将词袋转换为特征向量。例如，我们可以构造一个包含六个词语的词典：

序号	1	2	3	4	5	6
词语	铭铭	喜欢	打	篮球	也	乒乓球

我们将每个词语在文档中出现的次数按照词语序号排列起来，就得到这篇文档的词计数向量(term counting vector) n = (1, 2, 2, 1 , 1 , 1)。我们还可以对词计数向量进行归一化(即缩放向量的长度，使得所有元素的和为 1),得到词频向量 (term frequency vector) f=(1/8，1/4，1/4，1/8，1/8，1/8）。

通常我们并不要求词典包含文本中出现过的所有词语。如果文档中的某个词语并 没有在词典中出现，我们将其忽略即可。例如，如果使用下面这个只包含四个词语的

词典，这篇文档的词计数向量和词频向量就分别变为 $\boldsymbol{n}$ = (1, 2, 1, 1)与

$f=\left(\frac{1}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}\right)$。

序号	1	2	3	4
词语	铭铭	喜欢	篮球	乒乓球

在实际应用中，我们会使用一个公共的词典对语料库中的所有文档进行词频统计。我们以一个包含三篇文档的语料库为例：

文档 1：铭铭喜欢打篮球，也喜欢打乒乓球。

文档 2：铭铭去公园放风筝。

文档 3：铭铭的学校开设了人工智能课程。

首先，我们从语料库中提取所有出现过的词语，并形成一个词典。

序号	1	2	3	4
词语	铭铭	喜欢	打	篮球
序号	5	6	7	8
词语	也	乒乓球	去	公园
序号	9	10	11	12
词语	放	风筝	的	学校
序号	13	14	15	16
词语	开设	了	人工智能	课程

接下来，我们统计每篇文档中每个词语出现的次数，如图 7-2 所示。

	铭铭	喜欢	打	篮球	也	乒乓球	去	公园	放	风筝	的	学校	开设	了	人工智能	课程
铭铭喜欢打篮球，也喜欢打乒乓球。	1	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0
铭铭去公园放风筝。	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
铭铭的学校开设了人工智能课程。	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1

图 7-2 统计文档中出现的词

统计结果即是三篇文档的词计数向量：

n_1= (1, 2, 2, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

n_2= (1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0)

n_3= (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1)

词袋模型非常简单，但还需要与一些文本处理技术相搭配才能在应用中取得较好的效果。图 7-3 展示了利用词袋模型构建文本特征的基本流程。我们将在本节的后半部分对这些相关的技术进行简要的介绍。

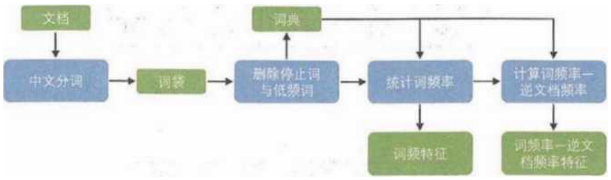


图 7-3 词袋模型应用的基本流程

中文分词

我们首先需要将句子中的词语分开，才能根据词语构建词袋。这个过程对于英文来讲是非常容易的. 我们只需要以空格和标点符号为依据，就可以将所有的词语分隔开来。但在中文文本中，所有的词语连接在一起，计算机并不知道一个字应该与其前后的字连成词语，还是应该自己形成一个词语。因此我们对文本构建词袋之前，需要先借助额外的手段将文本中的词语分隔开。这项技术称为中文的分词(word segmen- tation)。 中文分词方法大多基于匹配与统计学方法，我们在这里不做介绍。

停止词与低频词

在上面的例子中，我们看到词典中包含一些诸如“的”、“也”、“了”等词语。 这些词语是构成中文句子的基本字词，无论文档围绕什么主题，这些词语都不可避免地大量出现，但却对区分不同文档的主题毫无帮助。类似这样不携带任何主题信息的高频词称为停止词(stop word)在构建词典时，我们通常不会去除停止词。

在构建词典时，我们通常会去除出现次数极低的低频词。这类词语通常是一些不常用的专有名词。它们可能出现在特定的文章中(比如一篇采访稿中可能出现了 一位随机受访者的姓名)，但是并不能代表某一类主题。如果我们过度依赖这样的词语对文章的主题进行归类，那么就有可能出现过拟合的现象。另一方面，如果我们将这些低频词全部收录到词典中，就会大大增加词典的大小以及特征向量的维数，从而造成计算上的困难。因此，通常我们在收集完语料库中的所有词语之后， 会保留数千个或上万个常用词，而将低频词去掉。

词频率与逆文档频率

词频率与逆文档频率是反映一个词语对于一篇文档的重要性的两个指标。一个词语在一篇文档中出现的频率即为词频率（term frequency），它等于这个词语在这段文本中出现的次数与这段文本中词语的总数的商。我们记序号为 i 的词语在第 j 篇文档中出现的次数为 n_{ij}，那么第 j 篇文档的词语总数为 $n_j = \sum_{i=1}^{V} n_{ij}$，其中 V 是词典的大小。词语 i 在文档 j 中的词频率可以求得为 $\mathrm{tf}_{ij} = n_{ij}/n_j$。例如，第一篇文档中总共有四个词语，其中 1 号词“铭铭”在这篇文档中出现了 1 次，那么 1 号词在第一篇文档中的词频就是 1/4。

通常我们认为一个词语在一篇文档中出现的频率越高，这个词语对这篇文档的重要性就越大。例如，如果一段文本中大量出现“铭铭”这个词语，那么“铭铭”就很有可能是这篇文档的主要内容。但这种假设也有一定不合理之处，例如停止词在每篇文档中都会大量出现，但这些词语对一篇文档的重要性却是非常之低的。以前面出现的三篇文档为例，在第一篇文档中，1 号词“铭铭”和 4 号词“篮球”均只出现了 1 次，词频相同。但假如这三篇文档都出自铭铭的个人博客，那么“铭铭”一词就类似于停止词。无论这篇文档围绕体育运动还是围绕学校教育，“铭铭”一词都不可避免地出现，其重要性远低于“篮球”或“课程”这样可以区分不同主题的词语。这时，我们就需要借助逆文档频率（inverse document frequency）来修正每个词语在每篇文档中的重要性。

我们定义一个词语的文档频率（document frequency）为语料库中出现过这个词语的文档总数与语料库中所有文本的总数的商。如果语料库中总共有 D 篇文档，其中总共 D_i 篇文档中出现了第 i 个词语，那么第 i 个词语的文档频率即为 $\mathrm{df}_i = D_i/D$。而这个词语的逆文档频率即为文档频率的负对数，即 $\mathrm{idf}_i = \log(D/D_i)$。为了避免分母为 0 的情况，我们有时也将逆文档频率定义为 $\mathrm{idf}_i = \log(D/(1+D_i))$。逆文档频率同样刻画了词语在文本中的重要性，其值越高，重要性越大。

我们仍以前面的三篇文档为例。在去除“的”、“也”、“了”三个停止词后，每段文本都可以被表示为一个 10 维的词计数向量：

$n_1 = (1, 2, 2, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0)$

$n_2 = (1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0)$

$n_3 = (1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1)$

我们可以看出，1 号词“铭铭”在三篇文档中均有出现，则“铭铭”一词的逆文档频率为 $\log\left(\frac{3}{3}\right)=0$。“篮球”一词只在一篇文档中出现，因此它的逆文档频率为 $\log\left(\frac{3}{1}\right)\approx 0.47$。逆文档频率的计算值与我们的直观想法相符，即“铭铭”一词的重

要性要低于“篮球” 一词。

将一个词语在某篇文章中的词频率与该词的逆文档频率相乘，我们就可以得到这个词在这篇文章中的词频率--逆文档频率（tf-idf)。词频率-逆文档频率是对频率的一种修正，可以更好地突出文本中的重要信息。我们以将文档的词频向量中的频率值替换为 词频率-逆文档频率值，得到这篇文档的词频率-逆文档频率向量，作为文档的特征。

实验 7-1

在本实验中，我们将利用已经学习过的技术，将一个语料库中的文档转换为文本特征，作为后续主题挖掘任务的基础。

实验步骤：

1. 从教材工具包提供的语料库中任意选择几篇文档，阅读并概括文档的主题。
2. 利用教材提供的工具包，对语料库中的所有文档进行中文分词操作，去除停止词与低频词，并记录词典的大小。
3. 基于这个词典，计算出所有文档的词频向量。
4. 利用教材提供的工具包，针对词典计算每个文档词频率-逆文档频率向量。

7.3 高屋建瓴：发掘文本中潜在的主题

主题模型

主题模型（topic model)是描述语料库及其中潜在主题的一类数学模型。在主题 模型中，我们首先考虑的一个问题就是如何用数学语言来描述一个主题。在介绍词袋 模型时，我们知道文本中出现的词语可以反映文本的主题。那么如果我们可以搜集到 只包含某个单一主题“（例如图 7-4 中的教育主题）”的若干文档，并对其中词语的出现频率进行统计，那么统计的结果就可以作为这个主题的一种表示。

具体来说，如果词典的大小为 V，对其中每一个词语，我们统计其在所有文档中出现的总数 n_i，再除以文档中的总词语数 n，就可以得到对应的词频 f_i。我们再将所有的词频组合在一起，就可以得到一个维数为 V 的词频向 $t=(x_1, x_2, \cdots, x_v)$。 这个词频向量就是教育这个主题的一种数学上的表示。

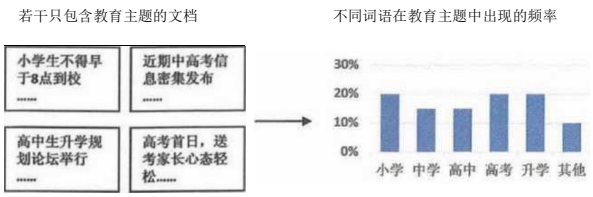


图 7-4 词频是主题的一种表示

词频统计的方法为我们提供了对主题进行建模的思路，但这种方法在实际操作中有其缺点。一方面，每一篇文档通常包含不止一个主题，单一主题的文档十分稀少。 另一方面，语料库中并没有关于文档主题的标注信息，即便存在单一主题的文档，我们也很难将其从海量的语料库中发掘出来。因此在实际操作中，我们必须借助额外的 技术来获取每个主题对应的词频向量。

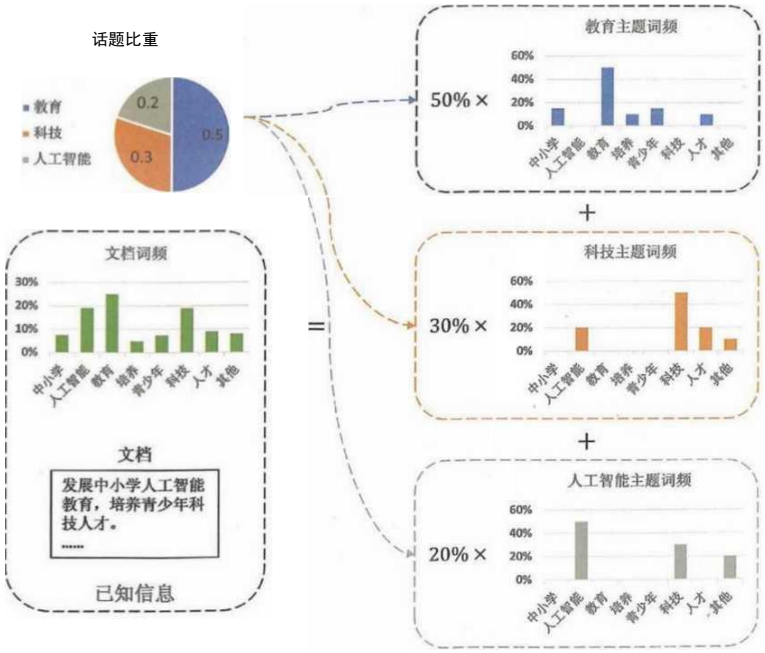


图 7-5 每篇文档的词频是由相关话题的词频按照比重混合而成

我们再来思考一下文档和主题之间的关系。我们知道，一篇文档通常会包含若干个主题，每个主题对应一个词频向量，比如图 7-5 中的例子中有“教育”“科技”和“人工智能”三个主题，这些主题的词频向量分别画在了图的右侧。那么这篇文档的词频向量和这些主题对应的词频向量之间又有什么样的关系呢？通常一篇文档所包含的各个主题的比重是不同的，在图 7-5 的例子中，“教育”话题的比重就比其它两个话题要大一些。在主题模型中，我们假设一篇文档的词频向量为其所包含的所有主题对应的词频向量的加权平均值，而每个主题对应的权重就代表了它在这篇文档中的比重。

具体来说，如果我们假设潜在的主题总共有 T 个（主题的个数通常是人工指定的。这与 K 均值算法中 K 的选择类似），每个主题对应于一个词频向量 $t_j=(x_{j1}, x_{j2}, \ldots, x_{jv})$，$1 \le j \le T$，在一篇特定的文档中，各个主题的比重分别为 $w_1, w_2, \ldots, w_T$。已知该文档的词频向量为 $d=(y_1, y_2, \ldots, y_v)$，则我们可以将文档词频、主题比重、 主题词频三者的关系表示为

$$d = w_1 t_1 + w_2 t_2 + \ldots + w_v t_v$$

其中 $w_i t_i$ 占为比重 w_i 与向量 t_i 的数量乘法。

借助矩阵的乘法运算，我们还可以将这个式子以更简洁的形式表示出来。首先我们将所有 T 个主题的词频向量排列成矩阵

$$\begin{bmatrix} — & t_1 & — \\ — & t_2 & — \\ \cdots & \cdots & \cdots \\ — & t_T & — \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1V} \\ x_{21} & x_{22} & \cdots & x_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{TV} \end{bmatrix}$$

随后，我们将所有主题比重排列成向量 $w = (w_1, w_2, \ldots, w_T)$。借助矩阵乘法运算，我们可以将公式(7-1) 简化表示为

$$d = wT \tag{7-2}$$

如果我们的语料库中总共有 D 篇文档，每篇文档的词频向量为 $d_K=(y_{K1}, y_{K2}, \ldots, y_{kv})$，$1 \le k \le D$，，每篇文档的主题比重向量为叫=$(w_{k1}, w_{k2}, \ldots, w_{kT})$，$1 \le k \le D$。我们可以将所有文档的词频向量与主题比重向量排列成矩阵

$$\begin{bmatrix} — & d_1 & — \\ — & d_2 & — \\ \cdots & \cdots & \cdots \\ — & d_D & — \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1V} \\ y_{21} & y_{22} & \cdots & y_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ y_{D1} & y_{D2} & \cdots & y_{DV} \end{bmatrix}$$

$$\begin{bmatrix} — & w_1 & — \\ — & w_2 & — \\ \cdots & \cdots & \cdots \\ — & w_D & — \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1T} \\ w_{21} & w_{22} & \cdots & w_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ w_{D1} & w_{D2} & \cdots & w_{DT} \end{bmatrix}$$

利用矩阵记号，我们可以将文档词频、主题比重、主题词频三者的关系表示为

$$D = WT \tag{7-3}$$

这个等式建立了语料库与潜在主题之间的关系，是主题模型的核心。

拓展阅读：矩阵乘法

下面我们通过一个例子来理解矩阵的乘法运算。假设我们的语料库中一共有 2 篇文档，词典中包含 2 个词语，我们考虑 3 个潜在的主题。

在这种情况下，主题词频矩阵 T 是一个 3 行 2 列的矩阵，每一行代表一个主题的词频向量，每一列代表了一个词语在所有主题中不同的词频主题比重矩阵 W 是一个 2 行 3 列的矩阵，每一行代表一篇文档中的主题比重。

矩阵相乘的结果仍然是一个矩阵，其行数等于第一个矩阵的行数，其列数等于第二个矩阵的列数。在我们的例子中，两个矩阵相乘的结果是一个 2 行 2 列的矩阵。

如图 7-6 所示，矩阵中每个元素的计算方法如下。我们从矩阵 W 中选取第 i 行，从矩阵 T 中选取第 j 列，计算这两个三维向量的内积，作为结果矩阵中第 i 行第 j 列的值。根据上述公式我们知道，这个值等于是第 i 篇文档中第 j 个词的词频。

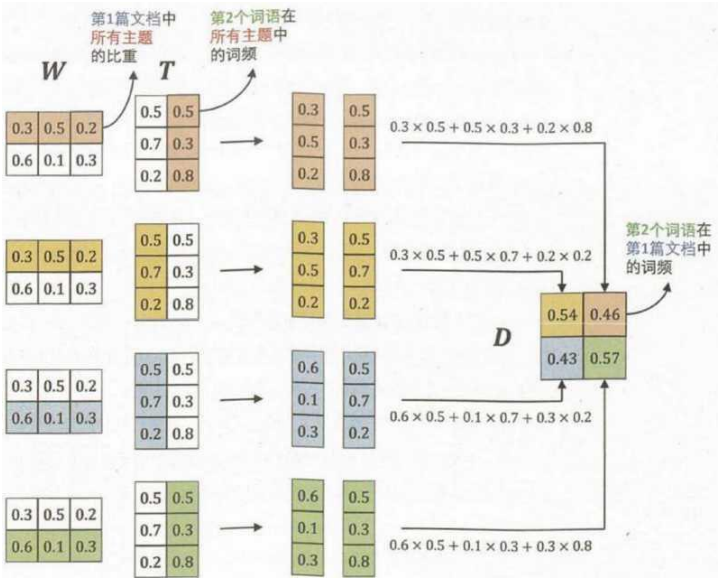


图 7-6 矩阵乘法示意图

潜在语义分析

通过主题模型，我们建立了语料库与其中潜在主题之间的关系。公式（7-3)实际上构建了一个方程组，其中等号左侧的文档词频矩阵可以通过统计语料库得到，是已知量，右侧的主题比重矩阵和主题词频矩阵都是未知量。通过求解这个方程组，我们就可以得到主题对应的词频向量，以及每篇文档包含的主题，就完成了对语料库中 潜在主题的发掘。这便是潜在语义分析技术。

但求解这个方程组也并不容易。我们注意到比更矩阵中有 DxT 个未知数，主题词矩阵中有 TxV 个未知数，而方程的个数则有 Dx V 个。一般来说，这个方程组中未知数的个数要少于方程的个数，这就意味着这个方程很有可能是无解的。之所以会出现这样的情况，就是因为我们的主题模型仅仅通过加权平均这样简单的方式建立起了语料库与主题之间的关系，不可避免地会引入误差。不过我们可以通过诸如非负矩 阵分解（non-negative matrix factorization)这样的方法，计算得到一组主题词频矩阵 T 和主题比重矩阵，使得公式（7-3)左右两边尽量接近。而由于非负矩阵分解涉 及较深的数学知识，我们这里不详细介绍。

主题词频矩阵 T 代表了语料库中所有潜在的主题，主题比重矩阵 W 包含了每一篇文档中各个主题的比重。在得到这两个解之后，我们就完成了对语料库中潜在主题的发掘，以及对语料库中每篇文档的概括与理解。

实验 7-2

住本实验中，我们将基于已有的文本特征，利用潜在语义分析技术对语料库进行主题发掘。

实验步骤：

1. 文档的词锁向量排列形成矩阵 D。
2. 设定主题数 T=10，利用教材提供的工具包，对矩阵 D 进行非负矩阵分解操作，得到词频矩阵 T 和主题比重矩阵 W。
3. 恨据词频距阵 T 列出每个主题的高频词，并尝试解释这些主题的含义。
4. 任选-两篇文章，阅读并概括主题，与解出来的主题比量向量进行比较。
5. 尝试使用词频率-逆文档频率代替矩阵 D 中的词频，重复上述实验，并比较结果。
6. 尝试更改主题数 T，重复上述实验，比较结果并思考：主题数过多或过少会有什么问题？

7.4 投其所好：基于主题的文本搜索与推荐

在日常生活与工作中，我们经常需要借助搜索引擎在互联网上寻找感兴趣的内容。传统搜索引擎通常基于关键词匹配技术。例如，如果我们希望搜索关于“食物中 的水分”的科普文章，搜索引擎就会去除停止词“中”和“的”，并利用关键词“食 物”和“水分”来查找匹配的文章。但是基于关键词匹配的搜索技术有两个明显的 缺点：

1. 近义词问题。例如我们以“开心”作为关键词，那么如“高兴”、“愉悦”等 近义词就会被忽略，但事实上它们的含义与“开心”是一样的。

2. 一词多义问题。例如“水分” 一词既可以表示物质上的水，也可以表示作品、 产品的虚假成分。如果我们以“水分”作为关键词，并希望搜寻科技类型的文章，仅通过关键词匹配，则一些例如打假类的文章就可能排在搜索结果的前列。

在学习了主题模型和潜在语义分析技术之后. 我们知道，每一篇文章都会包含若干个主题。如果我们在搜索过程中对文档的主题加以考察，就可以克服关键词的局限性。例如，如果我们在使用关键词“水分”进行搜索时，指定“科学”这个主题，那么和“虚假”相关的文章就很容易被过滤掉。

具体来说，我们首先会通过关键词匹配技术寻找包含“水分”的文章，作为候选的搜索结果。接下来，我们会利用通过海量文本分析得到的主题词频矩阵 T 对每一篇候选文档进行主题分解，得到每一篇文档所包含的各个主题的比重。随后，我们就可以筛选出与“科技”主题相关的文章作为最终结果返回给用户。

我们还可以借助潜在语义分析技术实现文章的个性化推荐。假如铭铭平时非常喜欢看新闻，新闻网站就可以在后台将铭铭看过的新闻搜集起来，并分析出这些新闻文档所包含的主题的比重向量。这些主题比重向量就代表了铭铭的偏好。如果铭铭平时喜欢看体育和科技类的新闻，那么体育和科技的比重可能就会很高。

随后，针对最新发布的新闻，我们也可以利用同样的技术分析出各个主题的比重。如果该比重向量与对应铭铭喜好的比重向量相近，那么这篇文章就很有可能是铭铭感兴趣的主题，新闻网站就可以将这篇文章推荐给铭铭。

7.5 本章小结

在本章中，我们由浅入深学习了词袋模型、主题模型，并最终利用潜在语义分析技术完成了文本分析与主题挖掘的任务。词袋模型只考虑词语在文本中出现的次数，忽略词语之间的顺序关系，是对文本建立的一个简单的数学模型。主题模型以词频代表主题，并假设文档的词频向址是文档所包含的所有主题的词频向量的加权平均。根 据主题模型的假设，我们可以列出一个关于语料库与其中潜在去题的方程，并利用非负矩阵分解的方法对其求解。

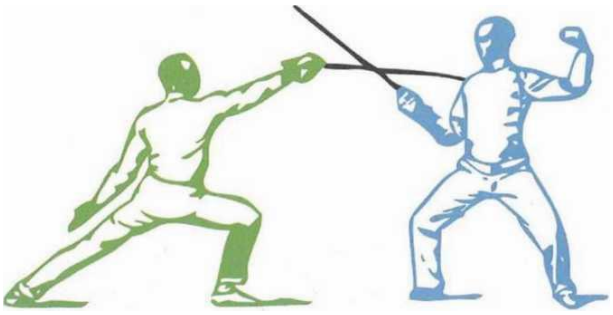
文本数据同时具有无监督与多主题或多类别的特点，而主题模型与潜在语义分析技术正是针对这两个特点而提出的一类重要的无监督机器学习方法。

神来之笔：创作图画

约翰翰尼斯•沸关尔
《基督在马大和马利亚家中》

汉•范•梅斡伦
《少年耶稣与长老》

二战期间，臭名昭著的纳粹德国元帅赫尔曼•戈坏(Hermann Goring)以两百幅名盛的代价从画家汉•范•梅赫伦(Han van Mee- geren)手里换得一幅荷兰名家约输尼斯•弗美尔(Johannes Vermeer) 的国宝级画作。战后，梅赫伦被控叛国通敌，等待他的将是无册的绞刑架，此时，梅赫伦坦陈他所出售的弗关尔的画作都是自己伪造的。在众人怀疑的目光下，梅排伦在狱中为自己的生金而画， 创作出了一幅“弗美尔味”十足的新作品《少年耶稣与长老》*(Je- lus amontr Uie Doctors)* 他的画不仅骗过了最有权势的元帅，也令专业的艺术评论家们挑不出蛛丝马迹从死，梅赫伦空前的仿画绝•技 名垂青史，而他在荷兰人心目中的形象也从一个叛国者变成了用秋 湛的绘画技术戏耍纳粹头目的民族英雄.



铭铭在读完梅赫伦的故事后非常激动，于是打电话给自己远在法国研习绘画的朋 友：“我想跟你学画画！”

“可是，我们远隔更洋，不能当面指导怎么办？”

“就没有别的办法了吗？”

“我想想，办法倒是有一个……每天你自己选定一个主题画一幅画传给我，我评 判一下画得好不好。如果觉得相好画还有差距. 就反馈给你进步的意见。然后你再根 据这些意见来完善绘画技术。这样，只要我们坚持‘创作一评判一反馈一完善’的流程，日复一日，你的创作水平就会越来越好。等到我再也挑不出毛病时，你就学成出师了。 ”

“所以我只要想办法能骗过你的眼睛——就像梅赫伦所做的一样——就成功了对不对？”

“没错，但小心我的鉴定水平也在进步哦！这其实也是一个有趣的对抗游戏。 ”

铭铭默默点点头，若有所思：这不是和生成对抗网络的原理有异曲同工之妙嘛！ 于是高兴地说：“你说得很有道理，我可以用相同的方法让机器也学会画画呢！”

那么，什么是生成对抗网络(generative adversarial network, GAN)呢？它为什么能赋予机器像画画这样的创造力呢？本章我们就来探讨一下如何用生成对抗网络来创 作出以假乱真的图片

8.1 九层之台，起于累土：数据空间和数据分布

“生成对抗网络”由“生成对抗”和“网络”三个词语构成。其中“生成 ”

是指它是一个生成模型（generative model），即它可以随机生成观测数据。举个例于，　假如给它的训练集是一些明星的照片，那么一个训练好的生成模型就可以“创作”　出全新的明星照片。生成对抗网络由生成网络（generative network)和判别网络(discriminative network)两个部分组成。其中生成网络就是用于生成数据，而判别网络则　是用来分辨数据是真还是假。以计算机自动创作图画的过程为例，生成网络在其中扮　演的角色就是艺术创作家（创作图画），而判别网络所扮演的角色是艺术鉴赏家（判　别图像是机器仿制的还是画家绘制的）。生成对抗网络的基本思想就是通过生成网络和判别网络之间的相互“对抗”来学习。那么，生成网络和判别网络具体是什么呢？　它们两者之间又是如何“对抗”的呢？为什么对抗会让生成质量越来越好呢？在回答这些问题之前，我们先回顾和介绍一下数据空间和数据分布这两个基本概念。

数据空间与数据分布

我们已经知道数据对人工智能系统的重要性，生成模型也不例外。假如，我们的目标是让计算机从无到有自动生成看起来像大牌明星的图片，就要提供大量的明星照　片供它学习参考。在生成模型眼里，这些照片数据组成一个整体，共同勾勒出明星们　的外观特点。生成模型不是要学习生成某个特定的明星的照片，而是要把握这些照片整体上的特点，生成有“明星范儿”的图片。那么，怎么刻画数据呢？这要引入数　据空间和数据分布的概念。

数据空间（dataspace)，顾名思义就是数据所在的空间，假定取的明星照片的分辨　率都是 128 x 128,回忆第三章的知识，每张照片就可以和一个 128 x 128 x3 的三阶张量　等同起来。此时的数据空间就是所有形状是 128 x128 x3 的张量的集合，或者说，就是这个分辨率下所有可能的图像构成的集合。在生成图像这个任务中，数据空间就是一　些图像的集合，所以我们也称之为图像空间。在图像空间中，每一张图片都是这个空间里的一个点。如图 8-1,我们看到，随便在图像空间里找一个点可能是一张没有意义的图片（图中绿色的点）。数据集里的图像也分布在这个空间里，我们，把这些特殊的点叫作数据点（图中黄色的点）。

知识链接：空间

在我们的经验中，我们生活的三维“空间”就是所有可能的位置点构成的整体。在数学中，空间（space) 的概念被推广，用来表示具有共性的元素组成的集合（set）。所以它不再特指二维空间或三维空间，而是可以表达更一般的概念。比如所有的数据组成数据空间，所有的图片组成图片空间，所有的向量组成向量空间，等等。

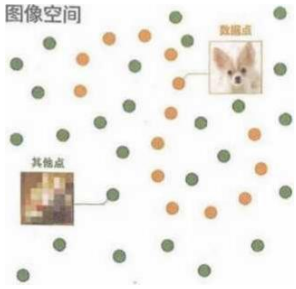


图 8-1 图像空间

数据点在数据空间中的分布情况是有一定规律的：空间中有些位置附近聚集的数 据点比较多，有些位置比较少或没有。这种数据在空间中的分布情况就称为数据分布 （data distribution)。在数学中，分布（distribution)是概率论分支里的一个基本概念。它与随机性紧密相关。我们通过下面的实验来认识随机性与分布的概念。

实验 8-1：认识随机性与分布

实验目的：对随机性有一个直观的感受。

实验步骤①：

生成 M 个随机数（以 M = 1 为例），并将随机数的值在图中画出，如图 8-2 所示。多次运行工具包中的样例代码，观察运行结果，体会生成数据的随机性。（注：每一行是一个样本）

要点：随机的直观感受就是每次运行的结果不一定一样。

结果参考：

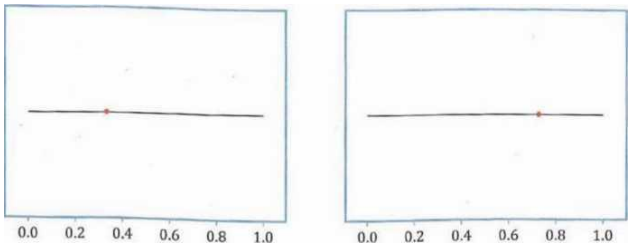


图 8-2 两次不同的运行结果（图中红色的点表示在区间[0,1]上随机生成的值）

实验步骤②：

画多个随机数(例如 M = 10, 100, 10000……)，观察输出的图中数值的分布特点。这里用直方图来刻画随机 性，如图 8-3 所示。

结果参考：

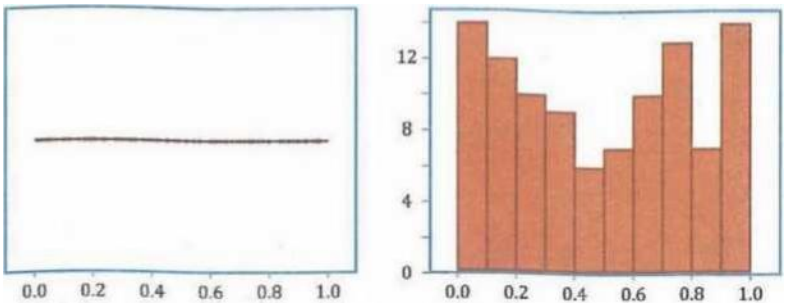


图 8-3 同时生成 100 个随机数连成的直方图

实验步骤③：

本步骤与步骤②类似，但要画正态分布(normal distribution)的直方图，结果如图 8-4 所示。

要点：随机性的体现方式多种多样，且它们的特点不完全相同。在生成图片等真实案例中，数据构成的分布比我们实验中所举的这些例子要复杂得多。

结果参考：

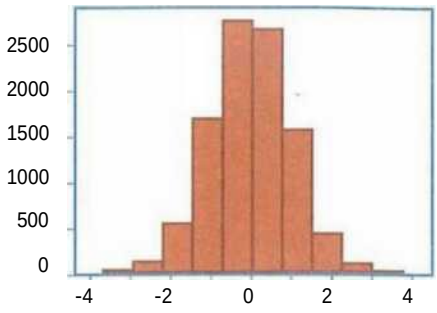


图 8-4 生成的随机数构成的正态分布的直方图

从这些实验中，我们认识了随机的概念，并了解了随机现象都服从着某种分布。通过直方图，我们可以直观地 看出不同分布的特点。所有明星的图片在图像空间中也构成了一个复杂的数据分布。这个分布不方便用直方图这样 的工具去直接刻画。那么怎么去描述数据分布呢？下面介绍的生成网络采用了一个巧妙的方法：把一个简单的、容 易把握的分布变成这个复杂的、难以把握的数据分布。这样就可以通过简单的分布间接地掌握复杂的数据分布。比 如说，这个简单的分布就可以选用前面实验中的正态分布。在生成对抗网络中，这个简单分布生成的样本所在的空 间我们就称为潜在空间(latent space)。

8.2 化腐朽为神奇的创作家：生成网络

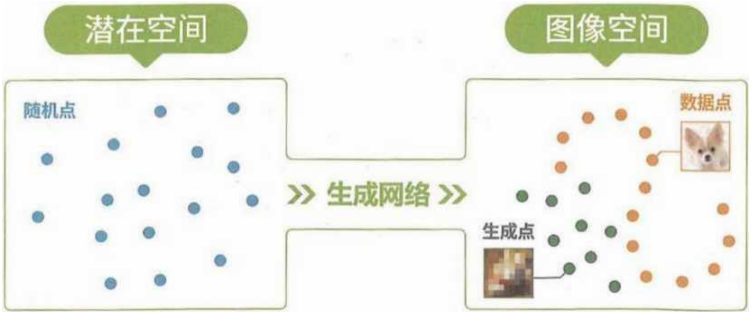


图 8-5 生成网络的工作任务图示（图像空间中的一个点代表了一张图片）

生成网络的职责是把随机点变成与数据集相似的图片。这些随机点是从一个潜在空间中随机抽取的。这就好比 艺术家把一个个原本简要抽象的艺术构思发展转化成了一张张具有复杂意象的画作。如图 8-5 所示，我们看到，生 成网络就是一个可以实现“点到点变换”的函数，它把潜在空间中的点变成图像空间中的点。生成网络生成的点就 叫作生成点。通过生成网络，潜在空间中的分布就可以变换为图像空间中的分布。我们将后者称为生成分布。有时 候生成网络也称为生成器（generator）。）。

实验 8-2：生成网络

实验目的：对生成网络的概念有直观的理解。

实验说明：生成网络就是用函数的映射来实现随机分布的变换。

实验步骤①：

假定生成器的数学含义可以用一个简单的线性函数 y = ax + b 来表示，则函数的图像如图 8-6 所示。生成一 批随机数，把生成的随机数代入函数中，可得到相应的输出。用直方图观察输入数据与输出结果的分布差异，如 图 8-7 所示。

要点：函数可以变换分布。

结果参考：

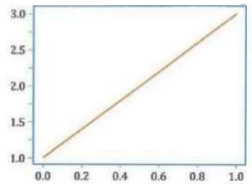


图 8-6 线性函数 y =2x + l 对应的图

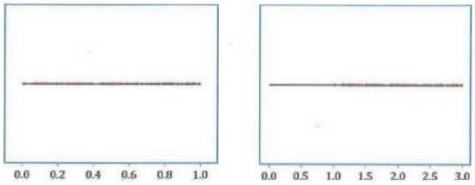


图 8-7 [0, 1] 上生成的随机数据代入函数 y=2x + 1 得到了介于[1，3]之间的结果

实验步骤②:

使用一个复杂的更换函数(函数图像如图 8-8 所示)，可以把函数结果的分布特点从均匀分布（uniform dis- tribution) 变换为正态分布，变换后的效果如图 8-9 所示。

要点：精心设计的变换函数可以把一种分布变成另一种分布。生成对抗网络就是要学习这样一种变换。

结果参考：

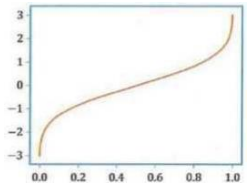


图 8-8 一个精心设计的变换函数

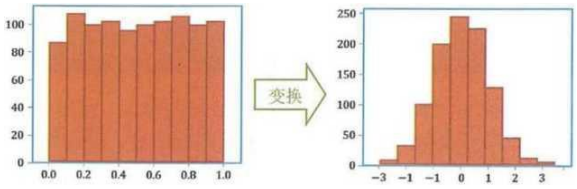


图 8-9 用一个精心设计的变换函数出把均匀分布（左）变成正态分布（右）

真实图像在图像空间中的分布情况是非常复杂的，简单的函数很难把这些随机的点恰好都变到真实图像所在的位置，所以实践中通常要利用深度神经网络。深度神经网络的强大的表达能力使得生成逼真的图片成为可能。然而， 一个随意设定的网络生成的通常是一些看起来毫无意义的图片。那么怎样训练生成网络才能让它生成有意义的图片呢？

我们回忆之前是怎么训练分类网络的。执行分类任务时，输入的是图片，输出的是类别。训练时，不论输入的图片是什么，分类网络都会为每个输入点找到一个确定的输出目标（如图 8-10 左图所示）。有了训练目标，我们可以通过减小与目标的差距而优化网络。

在生成网络中我们只提供了图片，除此之外，没有任何其他信息。所以潜在空间中的点在图像空间中没有一个确定的目标点，如图 8-10 右图所示。没有了直接对比的目标，那怎么优化生成网络呢？这时就需要生成对抗网络的另一个重要组件--判别网络。

图 8 - 10 分类网络与生成网络工作方式比较示意图

8.3 火眼金精的鉴赏家：判别网络

判别网络的任务是判断一张图片究竟是来自真实数据还是由生成网络所生成。在训练判别网络的过程中，通过不断给其输入两类不同的图片并为两类图片标注不同的数值以提高它的辨别能力。若输入的是真实数据中的图片， 标注数值 1,若输入的是当前生成网络生成的图片，标注数值 0。待判别网络训练好后，给它输入一张图片，如果它能确定这张图一定是由生成网络生成的，那么输出 0。反之，如果判别网络认为这张图一定来自真实数据，则输出 1。 有些情况下，判别网络认为一张图片既有可能由计算机生成也有可能来自真实数据，那么它就会输出图片是真实数 据的概率。

判别网络的输出结果用一个数值来指示空间中的一个点来自真实数据的可能性。输出数值 0 表示判别网络认为该图片一定是计算机自动生成的；输出数值 1 表示判别网络认为该图片一定来自真实数据；输出数值 0.5 则表示判别网络认为该图片有一半的可能是真实的，有一半的可能是计算机自动生成的。我们有时也称判别网络为判别器(discriminator)。

实验 8-3： 用判别网络分辨点的来源

实验目的：对判别器有一个直观的感受。

实验说明：判别器给出一个点来自真实数据的概率。

实验步骤①：

模拟数据分布与生成分布。假定数据点均匀分布在数轴上[2,5]之间，而生成点均匀分市在数轴上[1，3] 之间，画出图像（如图 8-11 所示），观察分布结果。

要点：这一步是为后续的实验步骤准备数据。

结果参考：

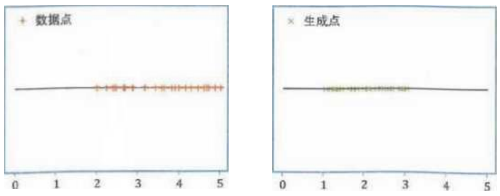


图 8-11 左图为真实数据的分布（在[2,5]之间）
右图为生成数据的分布（在[1,3]之间）

实验步骤②：

使用一个判别器对实验步骤①中的数据进行训练，生成图像（如图 8-12 所示），观察训练结果。要点：这一步骤是训练判别网络。

结果参考：

观察判别器的输出结果，能发现最优判别器的输出规律吗？实际上，最优判别器的输出表达式是：

$$\frac{p_{real}}{p_{real}+p_{fake}}$$

其中p_{real}与p_{fake}分别表示真实数据的概率密度（probability density）与生成数据的概率密度。

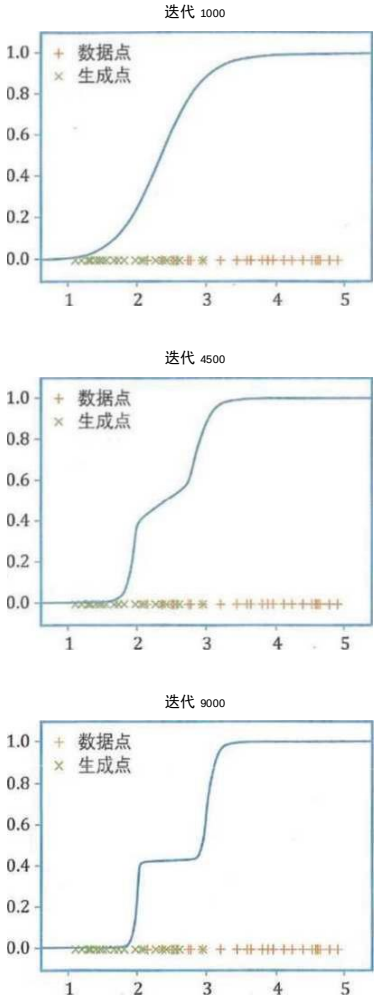


图 8- 12 训练过程中判别器的输出（三幅图分别是迭代 1000 次、4500 次、9000 次后的结果）

8.4 在对抗中合作与进步：生成对抗网络

生成对抗网络由生成网络和判别网络两部分组成。它们两者之间既相互协作又互相对抗。说它们相互协作，是因为它们是相互作用、相互扶持的。判别网络要想把真实图片与生成器生成的图片尽可能分清楚，就需要同时获得这两类图片。而生成网络要想生成与真实图片近似的图片，更需要依赖判别网络输出的反馈信息。说它们互相对抗，是因为判别网络的目标是慧眼识珠、明察秋毫，拒绝让生成网络生成的图片混入真实图片的行列之中。而生成网络的目的则是要尽可能地生成与真实图片类似的图像，从而让判别网络在判断时捉摸不透，达到以假乱真的效果。那么两者之间具体是怎样交互的呢？

生成对抗网络的训练包含两个交替进行的阶段，一个是固定生成网络用来训练判别网络，另一个是固定判别网络用来训练生成网络。

固定生成网络，训练判别网络

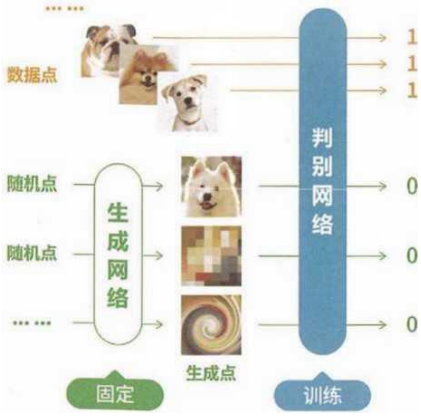


图 8-13 用固定生成网络训练判别器的工作模式示意图

图 8-13 展示了固定生成网络、训练判别网络的阶段。首先，我们生成一定数量的随机点，并利用生成网络将这些随机点变为生成图片。我们再找一定数量的真实图片，与生成图片组成一个二分类的数据集。分类的目标就是分辨图片是生成的还是来自数据集。在这个小数据集上，我们就可以训练一个判别网络，使之对真实图片的预 测接近 1，而对生成图片的预测接近 0，从而赋予其区分真实图片与生成图片的能力。

固定判别网络，训练生成网络

图 8-14 展示了固定判别网络、训练生成网络的阶段。在这个阶段中，我们持续地在潜在空间中生成随机点，并通过生成网络将这些随机点变换为生成图片。接下来我们将这些生成出来的图片输入到判别器中，并得到“该图片为真实图片”的概率。更重要的是，判别网络也会给生成网络提出如何提高判别输出概率的反馈信息。生成 网络利用得到的反馈信息来调整网络的参数，使得生成出来的作品能在判别网络中获 得更高的分数，经过一定量的训练之后，生成网络就可以输出更接近真实图片的生成图片。

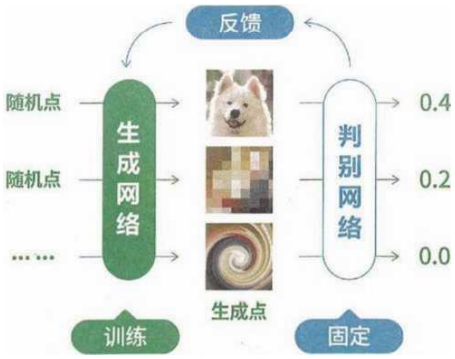


图 8-14 用固定判别网络训练生成器的工作模式示意图

•思考与讨论•

在学习了生成对抗网络的基本工作方式后，试想一下，在训练过程中随着生成器与判别器两者工作的动态交替，生成网络的图像创作水平能提高吗？生成网络生成的图片在判别器中得到的分数最后会趋近于 1 吗？

对抗过程演示

我们用一个虚构的简单例子来演示生成网络和判别网络之间对抗的动态过程。假定数据均匀分布在区间[3,5]中一开始生成网络只能生成区间[0，1]之间的 数。我们把它们在空间的分布情况用图 8-15 表示。

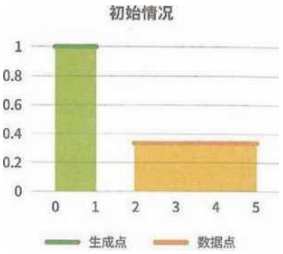


图 8-15 动态学习初始时，生成点与数据点的分布情况

图中，横轴代表数据点和生成点所在的空间。纵轴的数字是概率值。绿色和橙色的线分别表示生成点和数据点在空间里出现的概率，值越大表示出现的机会越大。所有值出现的机会的总和是 1,所以每条线下面的矩形的总面积都是 1。生成点的空间范 围小，在[0，1]之间，所以绿色的矩形又瘦又高；数据点的空间范围大，在[2,5] 之间，因此每个点的出现机会平摊下来就相对较小，橙色区域是一个又低又宽的矩形。

动态学习开始后，第一轮的运行结果如图 8-16 所示。

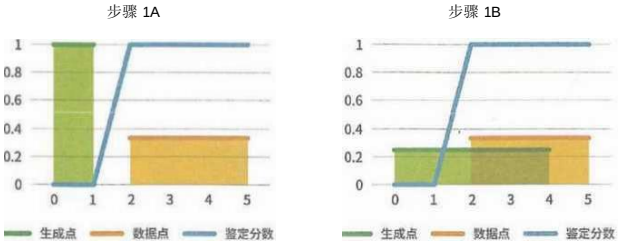
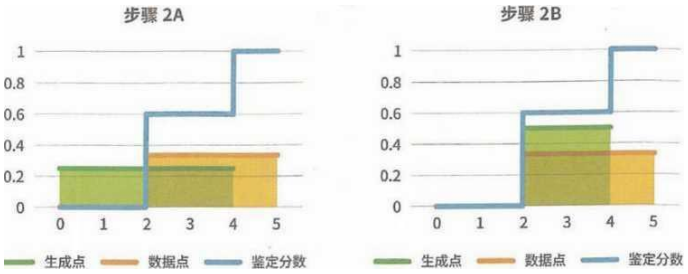


图 8-16 左图为判别网络经过训练得到每个点的判别结果，右图为生成网络生成范围调整 后的示意图

步骤 1A （如图 8-16 中的左图所示）是对判别网络进行训练后的结果示意图。图中，蓝色的线（鉴定分数）表示判别网络的输出分数。我们观察到生成点（图中绿色的线段）所在的区域离数据点（图中橙色的线段）所在的区域较远，这代表判别网络能清楚地分辨哪些点是生成器生成的（对应的输出得分为 0)，哪些点是来自真实数据的（对应的输出得分为 1)。

生成网络想追求更高的判别得分，于是采取了扩大生成范围的调整策略。通过步骤 1B（如图 8-16 中的右图所示）的调整，生成网络的输出范围和分数为 1 的区域 已经有一部分重合，说明生成器从判别器中得到了有用的反馈信息，并实现了一定的优化。

在第一轮运行过后，判别网络需要在第二轮运行中实现自我更新，以准确评估“对手”——生成网络的最新水平（如图 8-17 中的左图所示）。



图为 生成网络对生成点进行区同调整后的结果示意图

从图 8-17 的左图中可以看出，判别器对分布在区间[2, 4]之间的点所打的分值接近于 0. 6,表明判别器认为在这个区域中的点有约六成的可能来自真实数据集。此时生成网络得到反馈后会“战略性”地转移得分太低的点所在的区间范围（即空 间范围[0, 2]上的点），将点悉数置于区间[2, 4]内（如图 8-17 中的右图所示），扩大自己在[2, 4]之间的优势。从图中可以看到，在这个[2, 4]区间上生 成点出现的概率已经高于数据点。

•思考与讨论•

从图 8-17 中，我们观察到生成网络生成的点已全部包含于数据点的空间范围内。如果我们生成图片，那么现在所生成得到的图片都已经十分近 似真实图片了。唯一的缺陷就是对空间上［4, 5］之间的数据生成器还无 法生成，那么算法能解决这个问题吗？

在上一轮运行过后，判别网络再次自我更新，降低对区[2, 4]上的点的打分（如图 8-18 中的左图所示）。而生成器从判别网络中得知判别器对于区间[4, 5]之

间的点打分更高，所以生成网络将继续优化，优化后的结果如图 8-18 中的右图所示。 此时我们看到生成点所处的区域已经和数据点所在区域完美重合了。这时，大家不免会担心，如果继续运行下去，系统会不会把这个美好的状态打破从而变得更差呢？

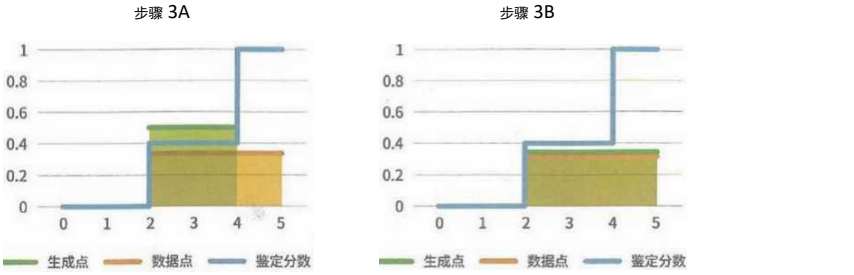


图 8-18 左图为判别网络更新后的数据结果关系示意图， 右图为生成网络优化后的结果示意图

在上一轮运行过后，判别网络对各区间内的点的输出结果都变成了 0.5（如图 8- 19 中的左图所示）。这意味着判别网络完全不能判断区间内的任何一个点究竟是生成器生成的还是来自真实数据的。而对于生成网络来说，它也没有改变的必要，因为生成点区间与数据点区间上各点的分数都一样，生成器认为自己生成的图片已经与真实图片没有区别了（如图 8-19 中的右图所示）。所以整个生成对抗网络达到了一个稳定状态。

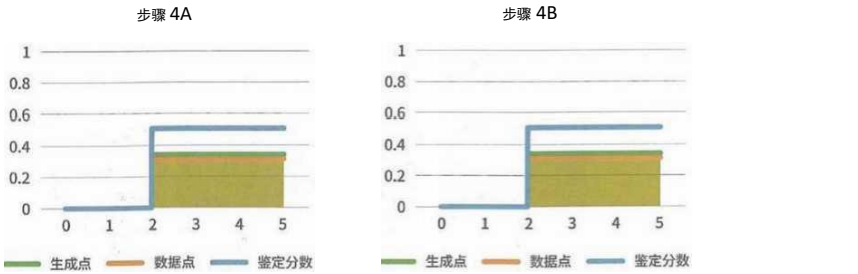


图 8- 19 左图为判别网络更新后的数据结果示意图，右图为生成网络未做改进进入稳定均衡状态

通过这样一个动态学习过程的演示，我们了解到在生成对抗网络中，生成网络与判别网络动态协作、相互竞争，最终达到一个理想的稳定状态。

实验 8-4：用生成对抗网络生成明星图片实验目的：生成看起来真实的明星图片。

实验说明：训练一个生成对抗网络。

要点：上手训练真实由片，对生成过程有一个直观的接触。 结果参考：

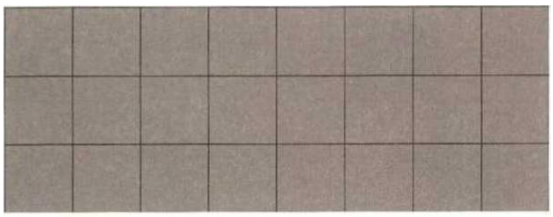


图 8-20 迭代 1，初始时生成的是简单灰色

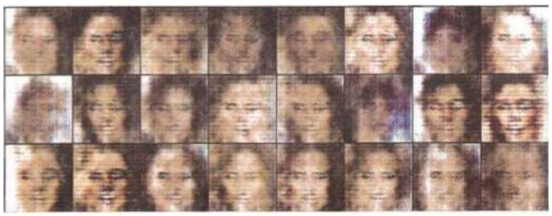


图 8-21 迭代 1000,， 生成的图像能初步看到人脸的雏形

图 8-22 迭代 3000，，生成的图像朝着正确的方向迈进

图 8-23 迭代 5000，部分生成图像已经有了较好的视觉效果

8.5 得心应手地创作：条件生成对抗网络

生成对抗网络虽然可以生成像真的图片，但是它牛成的内容却是随机的。如何能生成具有指定属性的图片，比如戴眼镜的明星呢？条件生成对抗网络(conditionnal generative adversarial network)解决了这个问题。它可以生成符合给定条件的图片。虽 然还没有进入实用阶段，但这种技术已经展示出了很大的应用前景。我们用图片展示两个可能帮助我们的例子：

•从侧脸到正脸：帮助识别罪犯

图 8-24 用条件生成对抗网络把侧脸变为正脸

在这个例子里，条件就是一个人的侧脸照片，而生成的目标则是同一个人的正脸图片(图片来自论文 “ Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis”) 。

•从年轻到年长：帮助寻找失散儿童

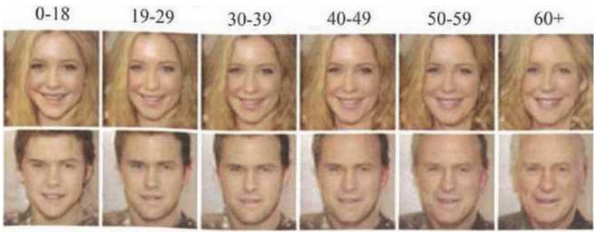


图 8-25 用条件生成对抗网络生成不同年龄的照片

在这个例子中，条件是给定年龄的入脸照片，而它的生成目标则是同一人在其他年龄的照片（图片来自论文“Face Aging with Conditional Generative Adversarial Nel- works”）。

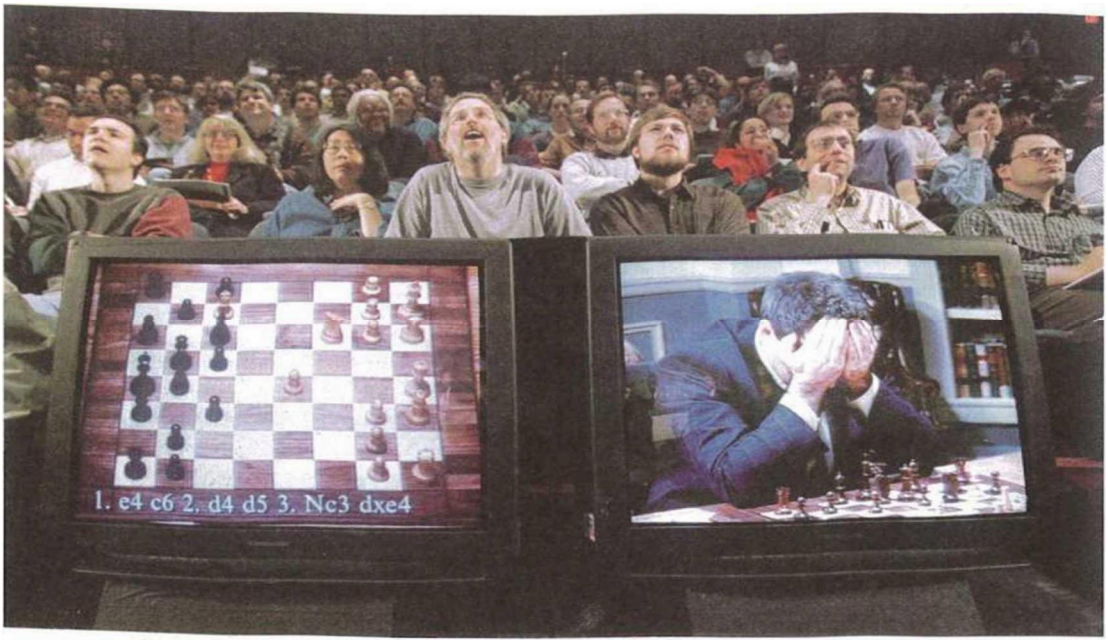
8.6 本章小结

本章为大家介绍了怎样借助生成对抗网络让计算机自动创作逼真的图像。在生成对抗网络的训练过程中，生成网络与判别网络相互协作、互相对抗，最终达到理想的均衡状态和近似于真实图像的生成效果。生成器为判别器提供了训练样本，判别器则为生成器提供了具体的优化目标。生成器的优化目的是以假乱真，用生成的图像骗过判别器的识别，判别器则要尽可能地分辨出输入的样本是真还是假。

生成模型是一类应用颇为广泛的人工智能模型。本章介绍的生成对抗网络是生成模型中的一个典型例子。一方面它应用在图像领域可以生成质量较好的图片，另一方面，对抗的思想也在其他任务中得到广泛的体现和应用。

第九章 运筹帷幄：围棋高手

2016 年，人工智能程序阿尔法狗（AlphaGo) 横空出世，横扫顶尖人类职业棋手，使得被称为人类智慧后堡垒的围棋，也最终被人工智能程序攻破。



故事还要从计算机下棋软件说起。在 1997 年时，IBM 的超级计算机“深蓝” (Deep Blue)击败了国际象棋世界冠军加里•卡斯帕罗夫，轰动世界。相比之下，围棋计算机程序的发展却非常缓慢，而其原因在于围棋棋盘下棋点多且局面多变，其复杂程度要远远高于国际象棋等其他棋类游戏，因此计算机想要在围棋中战胜职业棋手极为困难。自“深蓝”获胜以来，经过十多年的发展，棋力最高的围棋人工智能程序也只能达到业余棋手的水平，在不让子的情况下完全无法击败职业围棋选手。

阿尔法狗是谷歌旗下 Deep Mind 团队于 2014 年开始开发的人工智能围棋程序，它的出现为世界围棋史话抹上了浓重的一笔。2015 年 10 月，阿尔法狗击败欧洲围棋冠军樊麾，成为第一个无需让子即可击败围棋职业选手的计算机围棋程序。然而，不少评论员和人工智能专家仍然认为樊麾的棋力与围棋世界冠军相差甚远，阿尔法狗仍然无法击败顶尖围棋选手。2016 年 3 月，阿尔法狗与围棋世界冠军、职业九段棋手李世石展开了划时代的围棋人机大战，最终阿尔法狗以 4 比 1 的总比分获胜，再次震惊世界。2017 年 5 月，在中国乌镇围棋峰会上，升级版的阿尔法狗与排名世界第一的围棋世界冠军柯洁对战，以 3 比 0 的总比分完胜。至此，围棋界公认人工智能程序阿尔法狗的棋力已经超过人类职业围棋选手的顶尖水平。

对于阿尔法狗的大获成功，强化学习(reinforcement learning)功不可没，这也使得强化学习成为当前人工智能研究中的一个热点。让我们一起跟随阿尔法狗去探寻强

化学习的神奇之处，看看如何让计算机像人一样运筹帷幄，成为一名无师自通的围棋 高手吧。

9.1 初窥门径:阿尔法狗的走棋网络

正如我们之前所说，阿尔法狗是一个人工智能围棋程序，其功能便是在与人对弈的过程中，阿尔法狗给出当前局面(situation)应该在哪落子，以期最终在对弈中获胜。这个面对当前局面做出落子决第的过程由阿尔法狗中的走棋网络负责。走棋网络又被称为策略网络(policy network)，该网络接受当前棋盘局面作为输入，并输出在当前局面下选择每个位置的落子概率。

监督学习策略网络

阿尔法狗首先通过监督学习的方式训练了一个策略网络,被称为监督学习策略网络。监督学习策略网络使用了深度卷积神经网络来实现这一部分的功能，网络的输入不仅是当 前局面的落子状态，也加入了许多人为构造的特征，比如围棋中的气、目、空等。

对于该策略网络的训练，阿尔法狗团队从在线围棋对战平台 KGS 上获取了 16 万局人类棋手的对弈棋谱，并从中采样了 3000 万个样本作为训练样本。对于每个样本,包含当前棋局局面状态 s 以及人类落子方案 a，记为（s，a），将这 3000 万个训练样 本通过监督学习的方式训练该策略网络.从而得到了监督学习策略网络。实际上监督学习策略网络已经可以模拟人类棋手的风格进行落子

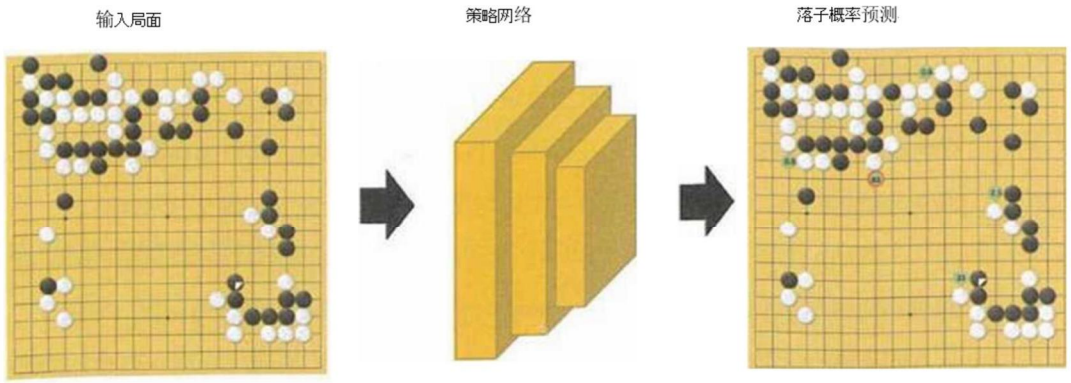


图 9-1 策略网络示意图

虽然已经可以模仿人类下棋，但实际上这种训练方法存在很大问题，因为 KGS 上的棋手水平高低不一，并非每个样本都是好的落子方案。而且顶尖棋手对局棋谱很少，拿这些样本来训练网络远远不够。确实，经过这样的训练，策略网络的棋力仍处于业余棋手的水平，完全无法与人类顶尖棋手过招。

分析其原因，除了棋手水平高低不一外，另一个重要的原因是训练样本(s，a）中只关心了棋手是怎么下的，而并没有关心最终结局是输了还是赢了，正所谓好的棋也学了，差的棋也学了，这样怎么能提高自己的棋力呢？

因此，为了进一步提高策略网络的棋力，阿尔法狗使用了强化学习技术。通过引入强化学习技术，阿尔法狗可以通过自我对弈左右互搏来提升自身棋力，从而变得更强。那么，到底什么是强化学习呢？

强化学习基本概念

强化学习与监督学习和无监督学习一样，同为机器学习算法中的一种。强化学习与监督学习最主要的区别就在于其收到的反馈是评估性的而不是指导性的。监督学习给出的指导性反馈将会通过监督信号告知学习者应该做出什么样的行为以获取更高的收益。而强化学习的评估性反馈意味着该学习系统只会告诉学习者当前的做法是好的还是坏的，所以学习者必须在多次尝试之后才能发现哪些行为会得到更高的回报；而且当前的行为不仅影响此时的回报，往往还会彰响后续的回报。

强化学习使得计算机能够像人一样通过完全自主学习来掌握一项技能，具有实现真正人工智能的潜力。而阿尔法狗也正是利用了这个自主学习的思想，通过自己跟自己的对弈来提升棋力。

主体与环境之间的交互

在了解强化学习的基本概念后，我们再来看看学习中强化有哪些基本的要素。

在强化学习中，我们称负责做出决策的实体为主体（agent),比如会下棋的阿尔 法狗、无人车、人等等。主体存在于环境(environment)中，其行为作用于环境，并接受环境的反馈。强化学习就是研究主体与环境之间的交互（如图 9-2 所示）。

主体通过发出特定的动作（action）来改变环境目前的状态（stale)，环境的状态 改变后将会返回给主体一个观察（observation)，同时返回给主体一个回报（reward)，而此时主体可以根据返回的信息发出新的动作，继续与环境进行交互。更通俗的理解，主体通过动作作用于环境后，环境的好与坏就通过回报反馈给主体。

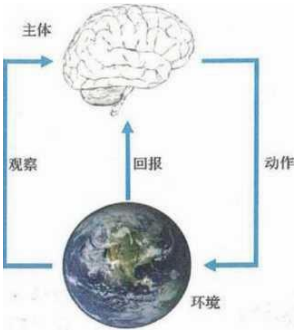


图 9-2 主题与环境之间的交换

在围棋中，阿尔法狗就是一个主体，其面对的棋局局面可以认为是目前的环境状态，阿尔法狗面对当前棋局状态，选择在哪里落子可以认为是阿尔法狗发出的一个动作，阿尔法狗观察到的落子后棋局的新状态即是环境返回的一个观察，而最终阿尔法狗是否赢得比赛就是环境反馈给阿尔法狗的回报。

策略与强化学习目的

环境的状态集合与主体的动作集合之间存在映射关系，即主体观察到环境的某个状态时，需要发出什么样的动作。更普遍地讲，在每个状态下，主体发出不同动作的概率往往是不一样的。策略（policy)指的是主体的行为，是一个从状态集合到动作集合的映射。比如在围棋中，环境的状态集合由所有可能的棋局局面组成，主体的动作集合就是阿尔法狗可以采取的所有符合规则的落子方案，而策略就是阿尔法狗的行为，即面对不同局面时阿尔法狗选择的下棋方案。

强化学习的目的就是找到一个最佳的策略，从而使得主体发出一系列的动作后，收到的累积回报最多。也就是说，我们需要通过强化学习让阿尔法狗找到一个最佳的策略，基于此策略选择落子，阿尔法狗可以在多次对弈中尽可能多地取得胜利。

强化学习策略网络

现在我们已经对强化学习有了一个基本的了解，那么强化学习到底是怎么训练的 呢？而且阿尔法狗又是怎么通过强化学习训练出了一个棋力更强的策略网络呢？

首先我们来了解一下强化学习的训练，该过程实际上是在主体与环境之间不断的 交互中完成的。主体将根据当前的策略不断地发出动作去改变环境的状态，而环境将

会根据状态的改变返回对应的回报，以表示当前动作是否对环境有利。而主体在收到环境的回报后，将根据回报的高低及时调整自己的策略，以期将来在新策略的指导下发出的动作能获得更高的回报。如此反复迭代，主体就可以根据环境的回报不断调整自己的策略，从而慢慢向最佳策略靠近。

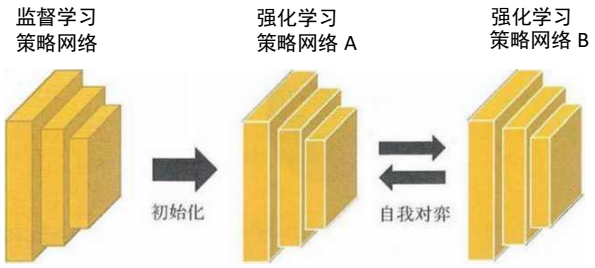


图 9-3 强化学习策略网络的

阿尔法狗使用了一种名为策略梯度的强化学习技术，训练了一个棋力更强的策略网络，名为强化学习策略网络。这个网络使用训练好的监督学习策略网络进行初始化，再通过不断的自我对弈，以最终胜棋为目标，迭代更新网络参数，从而改进策略来提高自己的获胜概率。每次自我对弈的双方是当前最新版本的阿尔法狗与随机选取的一个前几次迭代过程中的阿尔法狗。在每次对弈结束后，将根据当前版本阿尔法狗 在对弈中的胜负结果生成回报，最终获胜则反馈正的回报，否则为负的回报，而网络的参数将通过策略梯度技术朝着使回报最大化的方向变化。因此强化学习策略网络在 训练时的目标将不再是模拟人类棋手的风格进行落子，而是以最终赢棋为目标。

经过强化学习训练后的策略网络棋力大增，在与监督学习策略网络对弈时，已经可以取得80%的胜率。

在阿尔法狗策略网络的改进过程中，我们首次见识到了强化学习的强大之处，而在下一节中我们将继续学习强化学习是如何让阿尔法狗拥有了预测未来的能力。

9.2 远见卓识:阿尔法狗的大局观

虽然经过强化学习训练后的策略网络棋力大增，但是其缺点也很明显，即该版本的阿尔法狗只根据当前棋局局面就做出落子判断，这像是一个高手但只凭直觉在下棋，缺少所谓的“大局观”。所谓“手下一着子，心想三步棋”，顶尖的围棋高手在

对弈时都会在心中推演棋局的发展，从而帮助自己找到当前局面下更合适的落子方案。而缺少这种大局观的阿尔法狗显然仍无法匹敌顶尖围棋高手。

为了使阿尔法狗拥有这种大局观，Deep Mind 团队在阿尔法狗中引入了估值网络 (value network)，用于增强阿尔法狗对当前局面价值的判断，同时引入了蒙特卡罗树 搜索算法推演当前局面的发展，帮助阿尔法狗找到更好的落子方案。

估值网络

阿尔法狗中的估值网络用于量化评估当前围棋的局面，它使得阿尔法狗在对弈中无需走完全局即可快速预测当前局面的胜率。估值网络以棋盘当前局面作为输入，并预测阿尔法狗在当前局面下的胜率。

在训练估值网络时，阿尔法狗团队发现人类棋谱已经无法得出一个很好的估值函数，因此阿尔法狗再一次利用了强大的强化学习技术，即利用机器与机器对弈的方法 创造新的对局，从而产生足量的样本来训练估值网络。阿尔法狗通过使用训练好的强化学习策略网络进行自我博弈，从而产生了 3000 万个标注样本(s，z)，表示局面 s 下双方按照策略网络对弈后最终的胜负情况 z，并且每个样本都来自不同的一局棋，从而消除了样本之间的相关性。

基于自我对弈产生的样本训练出来的估值网络效果非凡. 至此阿尔法狗无需对弈到最后即可预测双方胜率, 这也使得阿尔法狗在有限时间内可以推演当前局面更多的可能性，从而找到更好的落子方案。

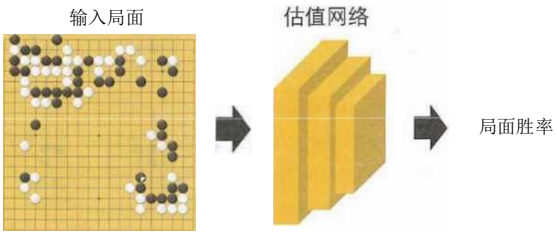


图 9-4 估值网络示意图

快速走子网络

为了加速棋局推演的速度，阿尔法狗中引入了快速走子网络，它可以说是一个轻量级的策略网络. 其落子效果虽不如策略网络，但速度却是策略网络的 1000 倍。快速走子的优点就是在之后进行蒙特卡罗树搜索时可以快速模拟更多的未来落子可能

性，从而帮助计算机更好地对当前局面进行评估。

蒙特卡罗树搜索

虽然有了估值网络可以判断当前局面的胜率，但这远远不够，阿尔法狗的大局观还是要从对当前局面的棋局推演中得到。因此阿尔法狗中引入了蒙特卡罗树搜索 (Monte Carlo tree search) 算法。

蒙特卡罗树搜索算法是一种通过随机推演建立一棵搜索树的启发式搜索过程，我们也可以将其看成是某种意义上的强化学习算法。在围棋中，蒙特卡罗树搜索算法会 从当前给定局面开始推演棋局，分别随机模拟双方落子，若干次后总有胜负，最终胜则回报为正，反之回报为负。之后该算法会反向沿着该对弈过程的落子方案一步步回溯，将路径上胜者所选择的落子方案分数提高，与此对应将败者的落子方案分数降低，所以之后遇到相同局面时选择胜者方案的概率就会增加。当不断重复以上步骤时，计算机将会试探很多种未来的落子可能性及其对应的胜负结果，那些好的落子方案的分 数也会不断提高，从而帮助计算机在当前局面下选择更有利于未来取胜的落子方案。

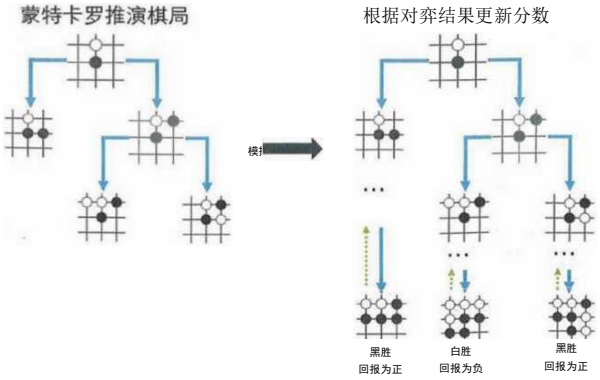


图 9-5 蒙特卡罗树搜索示意图

博采众长的阿尔法狗

阿尔法狗目前已经掌握了能输出每个位置落子概率的策略网络、速度超快的快速走子网络、能判断棋面价值的估值网络、能推演棋局发展的蒙特卡罗树搜索算法，那如何将这些部分整合起来形成一个博采众长的阿尔法狗呢？

答案就是在蒙特卡罗树搜索算法推演棋局时融合各个模块。

阿尔法狗面对当前棋局局面时，需要通过蒙特卡罗树搜索算法推演棋局的发展，即模拟对弈双方进行各种各样的落子试探。此时阿尔法狗在试探时每一步不再使用随机算法选择落子，而是根据在每个位置落子的预期收益来选择落子。此预期收益就融合了各个模块的功能，包括通过快速走子网络从当前局面开始多次模拟双方对弈直到分出胜负、策略网络对模拟中每个局面落子概率分布的估算、估值网络对模拟中每个局面价值的估算，并且该预期收益将在多次推演棋局发展中不断被 更新。

在每次模拟棋局对弈分出胜负后，此次落子方案中每一步落子动作的预期收益都 将被更新。因此经过多次落子试探后，我们对当前局面所有可选落子动作的预期收益将会有一个比较稳定的评估，从而帮助我们选择当前局面下更好的落子方案。实际上，阿尔法狗最终选择了在多次试探中当前局面访问次数最多的一个动作作为最终落子方案。

至此，阿尔法狗系统性地将策略网络、估值网络、快速走子网络整合到了蒙特卡罗树搜索算法中，博采众长，从而发挥出了巨大的威力. 最终战胜了顶尖围棋高手，一战成名。

9.3 成就非凡：阿尔法元

阿尔法狗一战成名后，Deep Mind 团队并没有沾沾自喜，而是继续推出了更强版 的人工智能围棋程序阿尔法元(AlphaGo Zero)。阿尔法元相比之前的阿尔法狗结构 更为简洁，且摒弃了人类棋谱的影响，完全通过自我博弈的强化学习算法训练自己，并且在与阿尔法狗的对弈中取得了100 比 0 的胜利。

强化学习算法在阿尔法元中发挥了更为重要的作用，可以说正是因为使用了强化学习，才使得阿尔法元无师自通，成为独孤求败的围棋天才。

阿尔法元概述

阿尔法元结构清晰而简洁，是强化学习算法的应用典范。阿尔法元在训练的开始就没有任何除规则以外的监督信号，并且只以棋盘当前局面作为网络输入，而不像阿尔法狗一样还使用其他的人工特征（例如气、目、空等）。此外，阿尔法元在模型上

只使用一个神经网络，该神经网络可以同时预测当前局面落子概率分布与局面胜率评估值，而不像之前版本一样分别使用策略网络与估值网络。

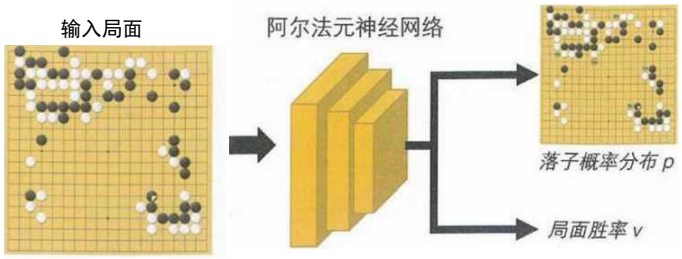


图 9-6 阿尔法元神经网络示意图

从强化学习的角度来看，阿尔法元使用了策略迭代的强化学习算法去更新神经网络的参数。简单来讲，策略迭代算法就是通过不断地交替进行策略评估和策略改进来完成强化学习的。接下来，我们将介绍一下阿尔法元是如何利用策略迭代的强化学习 技术来完成自我提升的。

阿尔法元的训练

阿尔法元的训练是通过不断的自我对弈来完成的。在每次对弈中，阿尔法元面对每一个局面时仍然执行蒙特卡罗树搜索算法推演棋局。与阿尔法狗一样，此时阿尔法元同样是根据当前局面每个动作的预期收益进行落子。但不同的是，阿尔法元可以通过一个神经网络同时预测出当前局面下的胜率评估值 v 和当前局面下的落子概率分布 P，从而用于更新对应动作的预期收益，且去除了快速走子网络，因此不再需要像阿尔法狗一样通过快速走子网络从当前局面开始模拟对弈直到分出胜负。

通过对当前局面进行多次蒙特卡罗树搜索算法推演棋局，阿尔法元最终可以搜索 出当前局面状态下的每个位置的落子概率分布实际上，阿尔法元也是通过进行大量蒙特卡罗树搜索试探，再统计当前局面下每个动作的选择次数来获得当前局面下的落子概率分布 π。而这个通过蒙特卡罗树搜索算法搜索出来的落子概率分布 π 往往比神经网络预测的落子概率分布 P 更优，即 π 可以作为 P 的目标值，因此蒙特卡罗树搜 索算法在阿尔法元的训练中实际上是一个策略改进过程。

阿尔法元在自我对弈中使用基于蒙特卡罗树搜索算法改进后的策略 π 进行落子， 并在自我对弈结束时统计胜负结果，将其作为策略迭代算法中的策略评估的标准，用于回溯更新网络参数。

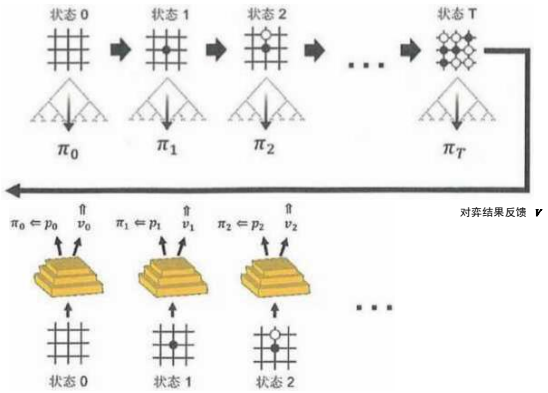


图 9-7 阿尔法元自我对弈训练 过程

在回溯更新神经网络参数时，阿尔法元将使得神经网络预测的落子概率分布 P 更 接近蒙特卡罗树搜索得到的落子概率 π，同时使得神经网络预测的局面胜负结果更接 近对弈最终的胜负结果。

阿尔法元可以说是完全使用了强化学习的架构，通过自我对弈来提升自己，并且 证明了在不使用人类棋谱训练的情况下，仍然可以打败结构更复杂且使用了人类棋谱 作为监督的阿尔法狗。阿尔法元同样也再一次向我们展示了强化学习算法的强大之处，它告诉我们在没有人类先验知识的情况下机器也可以在围棋这种高难度任务上战 胜人类。

9.4 本章小结

在本章中我们介绍了人工智能程序阿尔法狗背后的基本原理。我们了解到阿尔法狗由策略网络、估值网络、快速走子网络、蒙特卡罗树搜索四部分组成。伴随着阿尔 法狗原理的介绍，我们引入了强化学习算法。

强化学习是一个通用的决策框架，它使得计算机可以像人一样通过完全的自主学习提升自己，具备了实现通用人工智能的潜力。我们还介绍了强化学习的重要组成元 素以及强化学习的基本训练过程。通过介绍强化学习在阿尔法狗中的应用，我们看到了强化学习使阿尔法狗可以自我对弈，从而使策略网络变得更强，同样是强化学习使

得阿尔法狗可以训练出高效的估值网络，从而可以多次推演棋局，洞悉棋局发展。

最后，我们介绍了最强版人工智能围棋程序阿尔法元的基本原理，以及它与阿尔法狗的差别。阿尔法元为我们展现了一个更为简洁巧妙的强化学习架构，它使得阿尔法元实现了真正的无师自通.完全通过自我博弈成为围棋高手。

尽管我们已经看到了强化学习在围棋人工智能程序中发挥的巨大作用，但我们仍需明白，本章对强化学习的介绍也只是冰山一角。正如我们所说，强化学习是一种人工智能的通用框架，它在未来还有更多的可能性等着我们去探索。

后 记

当今世界，人工智能的浪潮正席卷全球。人工智能技术不再是象牙塔中的珍品， 它已经被大规模地应用于互联网、安防、医疗、金融、零售、文娱与教育等多个产 业，并实实在在地改变着我们的世界.让生产变得更加高效.让生活变得更加便捷。 国务院发布的《新一代人工智能发展规划》指出，人工智能已经成为国际竞争的焦点、经济发展的新引擎和社会建设的新机遇。

人工智能时代的建设和发展需要大批具有人工智能理念、国际视野和创新能力的人才。商汤科技作为中国领先的人工智能企业，秉承“坚持原创，让 AI 引领人类进 步”的使命，在不断进行科技创新的同时，也一直致力于推动人工智能教育的普及。 在 2017 年 11 月，商汤科技和华东师范大学慕课中心合作，组织力量与上海市六所重点中学的老师共同编写人工智能基础教材（高中版）。

在汤晓鸥教授和陈玉琨教授的指导下，商汤科技的编写团队（由林达华教授领衔）与上海巾車点中学的老师们（由田爱丽教授领衔）在几个月的编写过程中保持紧密沟通双方举行了多次联席研讨，对教材的每个方面，从篇章架构、概念组织到语言表述.都进行了充分的推敲和讨论。在大家的共同努力下，经历了多次修改打磨，书稿终于付梓。

在具体编写过程中，全书各章的撰写分工如下：

- 第一章：林达华
- 第二章：王鑫涛、陈向东
- 第三章：邵典、王若晖、陈恺、林勤、钱晋
- 第四章：余可、彭禹
- 第五章：张正夫、冻晨
- 第六辛：颜思捷、崔铭、金琼
- 第七章：沈岩涛、王若晖、孙时敏、敖培
- 第八章：李治中、李金杰
- 第九章：史少帅

与此同时，商汤科技的多位同事也义务为编写团队审读书稿，提出了很多宝贵的 建议和反馈。特别感谢以下同事为教材的编写和出版所付出的努力：乔宇、陈恺、王

佳琦、潘薪宇、尚海龙、金彦、邢晓菊、戴娟、颜深根、张富华、倪枫、周丹。特别感谢商汤设计部门的同事岳川、李莎、王艺璇为教材创作了大批生动活泼的图片。此外，也特别感谢商务印书馆的老师和编辑. 他们全程参与了教材编写过程的讨论，提 出了很多重要意见，并以很大的努力推动全书的付梓。

最后，我们感恩这个风云际会的大时代. 让我们有机会站在一起，见证人工智能这场新兴的科技革命如何改变世界. 让我们有机会为传承这场伟大变革的精神略尽绵 薄之力。

参考文献

第二章

[1] Fnnind, Yoav, and Robert E. Schapire. “Large inai-gin classificalion using the pemeptron algorithm. ” Machine leani- ing 37. 3 (1999)： 277 -296.

[2] Boser, Benihard E. , Isabelle M. Guyon, and Vladimir N. Vapnik. "A training algorithm for optimal margin classiG- ers. " Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992.

[3] Duda, Richard 0. , Peter E. Hart, and David C. Stork. Pattem classification. John Wiley & Sons, 2012.

[4] Christopher, M. Bishop. Pattern recognition and machine learning. Springer - Verlag New York, 2016.

[5] Viola, Paul, and Michael Jones. [44]Rapid object detection using a boosled cascade of simple features. ** Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Vol. 1. IEEE, 2001.

[6] Liu, Yun, el al. [u] Detecting cancer metastases on gigapixel pathology images. " arXiv preprint arXiv: 1703. 02442 (2017).

第三章

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoflrey E. Hinton. **linagenet classification with deep convolutional neural networks. ° Advances in neural iiilbruialion processing systems. 2012.

[2] LeCun, Yann, el ah **Gradient- based learning applied Lo document recognilion. " Proceedings of llie IEEE 86. 11 (1998) : 2278 -2324.

[3] Goodfellow, I. , Bengio, Y. , Courville, A. , & Bengio, Y. (2016) . Deep learning (Vol. 1) . Cambridge: MIT press.

[4] Zeiler, Matlliew D. , and Rob Fergus. [M] Visualizing and underslanding convolutional networks. ” European conference on computer vision. Springer, Cham, 2014.

[5] He, Kaiming. el al. "Deep residual learning for image recognition. [M] Proceedings of the IEEE conference on computer vision and pattem recognition. 2016.

[6] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing inlenial covariate sliifl. ” International conference on machine learning. 2015.

[7] C1FAR 10 数据集：https: //www. cs. toronto. edu/ ~ kriz/cifar. hlml.

第四章

[1] Tzanetakis, George, and Perry Cook. ** Musical genre classification of audio signals. ” IEEE Transactions on speech and audio processing 10. 5 (2002) : 293 -302.

[2] Sturm, Bob L ° The stale of the art ten years after a state of the art： Future research in music information retrieval." Journal of New Music Research 43. 2 (2014): 147 - 172.

[3] Goto, Masataka, el al. " RWC Music Database: Popular, Classical and Jazz Music Databases. ** ISMIR. Vol. 2. 2002.

[4] Oppenheim, Alan V · Discrete - time signal processing. Pearson Education India, 1999.

[5] Rabiner, Lawrence R. , and Biing - Hwang Juang. Fundamentals of speech recognition. Vol. 14 · Englewood Cliffs： PTR Prentice Hall, 1993.

第五章

[1] Zach, Christopher, Thomas Pock, and Horst Bischof. **A duality based approach for realtime TV-L 1 optical flow. ᴹJoint Pattem Recognition Symposium. Springer, Berlin, Heidelberg, 2007.

[2] Simonyan, Karen, and Andrew Zissenuan. ᵘTwo - stream convolutional networks for action recognition in videos. " Advances in neural iiifonnation processing systems, 2014.

[3] Wang, Limin, et al. **Temporal seginenl networks: Towards good practices for deep action recognition. " European Conference on Computer Vision. Springer, Cham, 2016.

[4] UCF 101 数据集 : http： //crcv. uc£ edu/data/UCFlOl. php.

第六章

[1] Viola, Paul, and Michael Jones. ᵈᵈRapid object deteclion using a boosted cascade of simple features. " Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Vol. I. IEEE, 2001.

[2] Ren, Shacxpng, el al. ᴹ Face alignment at 3000 fps via regressing local binaiy features. " Proceedings of the IEEE Con · ference on Computer Vision and Pattern Recognition. 2014.

[3] Sun, Yi, et al. **Deep learning face representation by joint idenlincation - verification. " Advances in neural information processing systems. 2014.

[4] Sun, Yi, Xiaogang Wang, and Xiaoou Tan 各 "Deep learning face representation from predicting 10 , 000 classes. " Proceedings of lhe IEEE Conference on Computer Vision and Pattern Recognition, 2014.

第七章

[1] Landauer, Thomas K. Lalenl semantic analysis. John Wiley & Sons, Inc, 2006.

[2] Hofmann, Thomas. ᵘProbabilistic latent semantic analysis. ᴹ Proceedings of lhe Fifteenth conference on Uncertainty in artificial intelligence. Morgaii Kaufmann Publishers Inc. , 1999.

[3] Blei, David M. , Andrew Y. Ng, and Michael I. Jonlan. ᵘLatent dirichlet allocation. " Journal of machine Learning research 3. Jan (2003): 993 - 1022.

[4] Lee, Daniel D. , and H. Sebastian Seung. ᵘAlgorilluns for non - negative matrix factorizalion. ᴹ Advances in neural iiilbmiation processing systems, 2001.

[5] Gaussier, Eric, and Cyril Goulte. **Relation between PLSA and NMF and implications. " Proceedings of lhe 28th annual international ACM SIGIR conference on Research and devclopmcnt in information retrieval. ACM. 2005.

第八章

[1] Goodfellow, Ian, et al. ᵘ Generative adversarial nets. '* Advances in neural infonnation processing systems, 2014.

[2] Huang, Rui, el al ・"Beyond face rotation: Global and local perception gan for photorealistic and idenHly preserving frontal view synthesis. arXiv preprint arXiv: 1704. 04086 (2017).

[3] Antipov, Grigory, Moez Baccouche, and Jean - Luc Dugelay ・"Face aging with conditional generative adversarial net ・ works. arXiv preprint aiXiv: 1702. 01983 (2017).

[4] Radford，Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional gen ・ erative adversarial networks. " arXiv preprint arXiv: 1511. 06434 (2015).

第九章

\ 1] Wang, Fei - Yue, et al. ** Where does AlphaCo go： From church - luring thesis to AlphaCo thesis and beyond. IEEE/CAA Journal of Automatica Siiiica 3. 2 (2016): 113 - 120.

[2] Chen, Jim X. **The evolution of computing： AlphaGo. ᴹ Computing in Science & Engineering 1&4 (2016) ： 4 - 7.

[3] Sullon, Richard S. , and Andrew G. Barto. Reinforcement learning： An introduction. Vol. 1. No. I. Cambridge: MIT press, 199&

[4] Kaelbling, Leslie Pack, Michael L. Littman, and Andrew W. M<x)re. **Reinforcement learning： A survey. " Joumal of artificial intelligence research 4 (1996) : 237 -285.

[5] Dowding, Keith. ** Model or metaphor? A critical review of the policy network approach. Political studies 43. 1 (1995)： 136 -158.

图书在版编目（C1P)数据

人工智能基础：高中版/汤晓鸥，陈玉琨主编. — 上海：华东师范大学出版社,2018

ISBN 978 - 7 - 5675 - 7561 - 5

I . ①人…II . ①汤…②陈…HL①人工智能一 高中一教材 IV.①G634.671

中国版本图书馆 CIP 数据核字（2018 ）第 050537 号

人工智能基础（高中版）

主 编 汤晓鸥陈玉琨

责任编辑 龚琬洁孙婷

纣而设计 李明轩

出版发行 华东师范大学出版社（上海市中山北路 3663 号 200062）

www.ecnupress.com.cn

商务印书馆（北京王府井大街 36 号 100710） www.cp.com.cn

印 刷 上海书刊印刷有限公司

开 本 889 x 1194 16 开

印 张 11.25

字 数 230 千字

版 次 2018 年 4 月第 1 版

印 次 2018 年 6 月第 4 次

书 号 ISBN 978-7-5675 -7561 -5/G.11009

定 价 35.00 元

出版人 于殿利王焰

（如发现本版图书有印订质址问题. 请寄回本社市场部调换或电话 021-62865537 联系）