HZ BOOKS
华章教育

智能科学与技术丛书

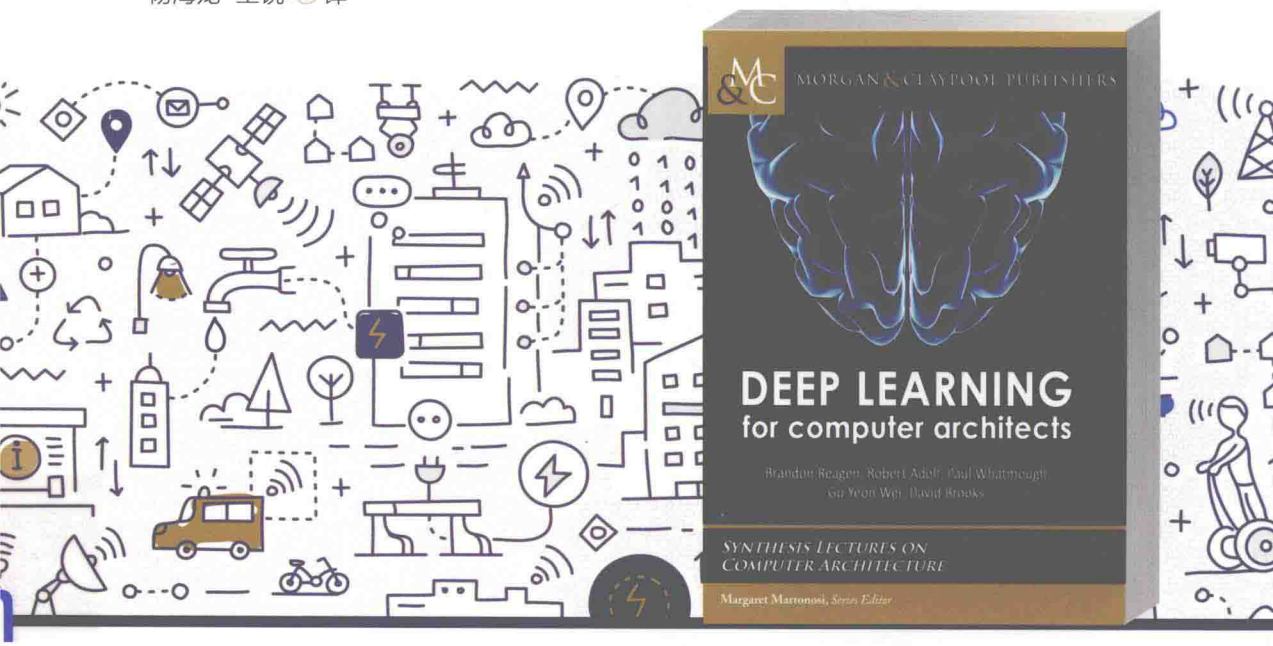
Deep Learning for Computer Architects

当计算机体系结构遇到深度学习

面向计算机体系结构设计师的深度学习概论

布兰登·里根（Brandon Reagen）
罗伯特·阿道夫（Robert Adolf）
[美] 保罗·沃特莫（Paul Whatmough） ◎ 著
古杨·魏（Gu-Yeon Wei）
大卫·布鲁克斯（David Brooks）

杨海龙 王锐 ◎ 译



机械工业出版社
China Machine Press

智能科学与技术丛书

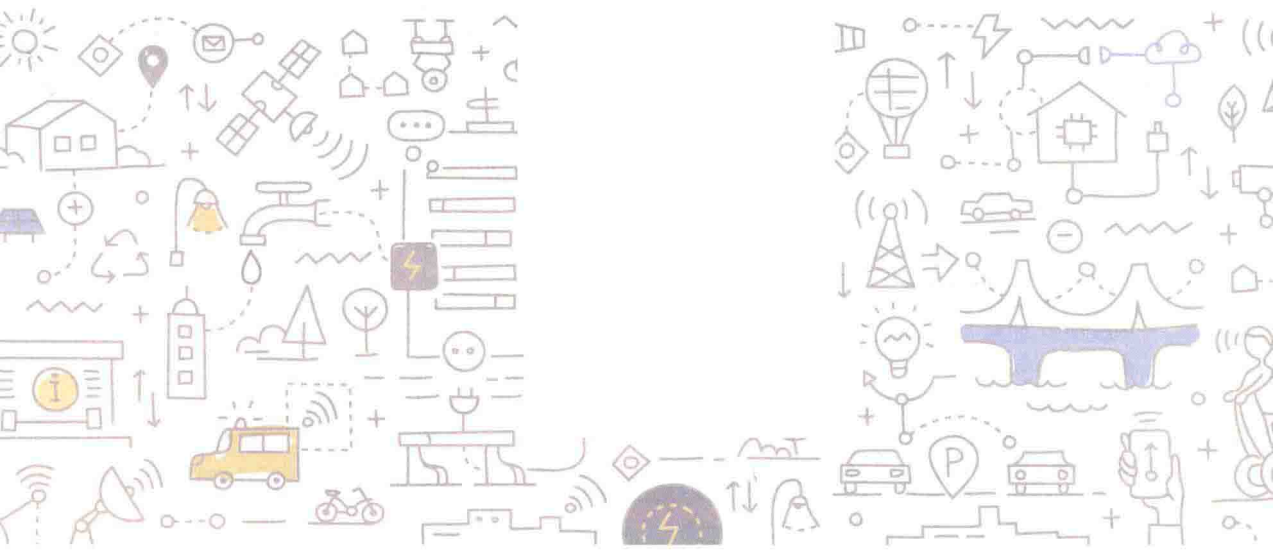
Deep Learning for Computer Architects

当计算机体系结构遇到深度学习

面向计算机体系结构设计师的深度学习概论

布兰登·里根（Brandon Reagen）
罗伯特·阿道夫（Robert Adolf）
[美] 保罗·沃特莫（Paul Whatmough） 著
古杨·魏（Gu-Yeon Wei）
大卫·布鲁克斯（David Brooks）

杨海龙 王锐 译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

当计算机体系结构遇到深度学习：面向计算机体系结构设计师的深度学习概论 /（美）布兰登·里根（Brandon Reagen）等著；杨海龙，王锐译 . —北京：机械工业出版社，2019.4
（智能科学与技术丛书）
书名原文：Deep Learning for Computer Architects

ISBN 978-7-111-62248-2

I. 当… II. ①布… ②杨… ③王… III. 机器学习－研究 IV. TP181

中国版本图书馆 CIP 数据核字（2019）第 050494 号

本书版权登记号：图字 01-2017-8756

本书是面向计算机体系结构设计师的深度学习入门读物。书中首先介绍机器学习的发展历程，并追踪深度学习技术的关键发展阶段。然后，回顾了代表性的工作负载，包括各种领域中常用的数据集和开创性的神经网络。接下来，详细介绍了颇受欢迎的深度学习工具，并展开介绍了如何使用工具与工作负载来表征和优化 DNN。本书的其余部分致力于介绍如何设计和优化用于机器学习的硬件和体系结构，并对近年来提出的各种优化方法进行重新梳理，以便进一步改进未来的设计。最后，本书回顾了该领域新近发表的研究文献并对其进行分类，帮助读者理解各种贡献的背景和意义。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：杨宴蕾　　责任校对：李秋荣
印　刷：北京文昌阁彩色印刷有限责任公司　　版　次：2019 年 4 月第 1 版第 1 次印刷
开　本：185mm×260mm 1/16　　印　张：9
书　号：ISBN 978-7-111-62248-2　　定　价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88379833　　投稿热线：（010）88379604
购书热线：（010）68326294　　读者信箱：hzjsj@hzbook.com

THE TRANSLATOR'S WORDS

译者序

近年来，人工智能特别是深度学习不断成为新闻头条出现在人们的视野中，特别是它在困难分类和回归问题上取得的巨大成功，例如图像分类、语音识别、自动翻译等领域，展现出了深度学习在真实应用中的颠覆性影响，并推动着计算机多个学科的重大变革。深度学习之所以取得如此显著的成绩，得益于海量数据集和高性能硬件的出现，特别是利用高性能硬件的体系结构特点，对深度神经网络模型进行软硬件联合设计和优化，将是未来深度学习领域出现重大突破的关键。随着深度学习体系结构成为未来计算机体系结构领域的研究重点，如何讲授相关的研究成果和进展，对推动该领域发展是很大的挑战。

由于深度学习的体系结构研究仍然处于迅猛发展的阶段，缺少针对该领域提纲挈领性的参考书籍，而本书的出现正好弥补了这部分空白。作者基于在深度学习和体系结构领域的长期研究和实践经验，全面介绍了深度学习的发展、评测和优化的最新研究成果，帮助读者掌握深度学习体系结构的精髓。本书大致可以分为四部分内容。首先，介绍了神经网络的基本知识和发展历程，从线性回归到感知器，以及当前最先进的深度神经网络。其次，主要针对当前流行的深度学习软件的设计异同进行了深入介绍，指导读者针对具体应用选择最正确的软件。同时，展示了 Fathom 深度学习评测集，该评测集能够帮助科研人员更好地评价研究成果的贡献。再次，对利用定制化硬件加速神经网络的体系结构进

行了探索，特别是针对 Minerva 加速器设计和优化框架，具体介绍了 Minerva 方法论以及如何设计实验在神经网络准确度、功耗、性能和硬件面积间进行取舍。最后，给出了神经网络论文中有关硬件研究的全面综述，并且提出了一种分类方法，帮助读者理解和对比不同的研究项目。总之，本书既可以作为高等院校高年级本科生、研究生以及受过计算机科学或工程训练的专业人士的教科书，也可以作为深度学习处理器设计专家的参考书。

本书作者均来自美国高等教育顶级学府哈佛大学，其中既包括从事第一线研究的博士生（Brandon Reagen 和 Robert Adolf），也包括资深研究人员和教授（Paul Whatmough、Gu-Yeon Wei 和 David Brooks）。

译者在翻译本书的过程中，正好有幸参与国内深度学习处理器相关的研究课题，在接到机械工业出版社华章公司的翻译邀请之时备感兴奋。一来本书在译者从事课题研究的过程中，对于快速了解整个领域的研究前沿起到了提纲挈领的作用，是从事本领域研究不可多得的优秀参考书；二来由于该领域研究发展迅猛，国内缺乏对该领域进行全面介绍的书籍，译者希望能在优秀参考书引进方面做出微薄的贡献，加速该领域知识在国内的传播。

书中有些术语目前还没有统一译法，所以在翻译过程中保留了其英文名称。由于时间和水平有限，译文中难免存在错误和不妥之处，恳请广大读者和同行不吝批评指正。本书在翻译过程中得到了北京航空航天大学计算机学院老师和学生的大力支持。另外，本书的出版还得到华章公司的大力帮助，在此对出版社同仁在排版和校对等环节的辛勤付出表示衷心的感谢。我们希望本书的出版对于国内深度学习体系结构的研究和人才培养起到促进的作用。

杨海龙　王　锐

PREFACE

前 言

本书旨在为具有计算体系结构、电路或者系统背景的研究者和设计师提供一份对神经网络的概述。引言部分（第 1 章）定义了关键词汇表，介绍了该项技术的历史和发展过程，并阐述了该领域需要额外硬件支持的原因。

接着，本书回顾了神经网络的基本知识，从线性回归到感知器，以及当前最先进的深度神经网络（第 2 章）。本书涵盖的范围和使用的语言使得任何读者都可以理解，并且本书的目的是让整个社区在对深度学习的认识上达成一致。虽然人们对该领域的研究兴趣激增，但仍有证据显示许多术语被混淆在一起，并且对该领域的理解也存在着差距。我们希望本书呈现的内容能够澄清对该领域的错误理解，并为非专家的读者提供一个统一的基础。

在回顾之后，将深入介绍工具、工作负载和表征。对于实践者，这可能是最有用的一章。该章首先综述当代神经网络和机器学习的软件包（例如，TensorFlow、Torch、Keras 和 Theano），并且解释这些软件包的设计选择和不同之处，从而指导读者针对自己的工作选择正确的工具。在第 3 章的后半部分，展示了一组常用的、开创性的工作负载，并且将其整合到了名为 Fathom 的评测集中[2]。然后，进一步将这些工作负载分为两类——数据集和模型，并且解释了为什么这些工作负载和数据集具有开创性以及该如何使用它们。这部分内容同样能够帮助神经网络论文的评审

人员更好地评价论文的贡献。通过更好地理解每个工作负载，我们可以更加深入地解释其背后的想法和贡献。伴随着评测集，是对工作负载在CPU 和 GPU 上的表征分析。

第 4 章构建在第 3 章的基础上，可能是那些探索利用定制化硬件加速神经网络的体系结构设计人员更感兴趣的部分。在本章中，回顾了 Minerva 加速器设计和优化框架[114]，并且详细介绍了如何将高级别的神经网络软件库与硬件 CAD 和模拟流糅合在一起来设计算法和硬件。本书特别关注 Minerva 方法论以及如何设计实验在神经网络准确度、功耗、性能和硬件面积间进行取舍。读完本章后，研究生应当可以掌握如何评价自行设计的加速器或者定制硬件优化的优劣。

在第 5 章中，给出了神经网络论文中相关硬件的全面调查和综述，并且提出了一种分类方法，帮助读者理解和对比不同的项目。本章主要关注过去 10 年的研究工作，并将论文按其所针对的计算栈层次（算法、软件、体系结构或者电路）以及优化类型（稀疏性、量化、计算近似和容错）进行分组。本综述主要关注机器学习、体系结构和电路领域的顶级会议，并尝试在本书出版时涵盖与本领域的体系结构设计师最相关的工作。但实际情况是有太多已发表的论文无法同时包含进本书。本书希望：这里的综述可以作为一个起点；分类提供一个顺序，供感兴趣的读者了解去哪里可以学习到一个具体主题的更多内容；对神经网络硬件支持的非正式讨论可以提供一种比较相关工作的方法。

最终，在本书的总结部分，通过列出仍待完成的工作，澄清了关于深度学习研究的硬件已经达到饱和点的谬论。尽管在这个主题上已经有大量的论文，但即便是在监督学习这个方面，也仍然距终点有很长的距离。本章阐明需要关注的领域，并简要概括了机器学习的其他领域。更

进一步，虽然对于机器学习社区，硬件大部分情况下只是一个服务产业，但是体系结构设计师确实应该开始思考如何利用当代机器学习来改善硬件设计。这个过程并不容易，因为它需要真正地理解这些方法，而不是实现已有的设计。但是如果说机器学习在过去的 10 年中教会了我们什么，那就是这些模型效果很好。计算机体系结构属于计算机科学中最少理论化的领域（几乎是完全基于经验和观察）。机器学习可能会在重新考虑如何设计硬件方面提供很多帮助，包括贝叶斯优化，并展示这些技术在硬件设计上带来的好处。

ABOUT THE AUTHORS

作者简介

布兰登·里根（Brandon Reagen）是哈佛大学的博士生。他于2012年获得马萨诸塞大学阿默斯特分校计算机系统工程和应用数学专业的学士学位，并获得了哈佛大学计算机科学专业的硕士学位。他的研究涉及计算机体系结构、VLSI和机器学习领域，他特别关注设计极其高效的硬件，以便能在所有计算平台上普遍部署机器学习模型。

罗伯特·阿道夫（Robert Adolf）是哈佛大学计算机体系结构的博士生。他于2005年从美国西北大学获得计算机科学专业学士学位，此后他就职于国防部，从事超级计算机基准测试和性能分析工作4年。2009年，他作为研究科学家加入太平洋西北国家实验室，领导一个团队在大规模多线程体系结构上构建大规模图形分析。他的研究兴趣是高性能软件的建模、分析和优化技术，目前主要关注深度学习算法。他的理念是，将统计方法、代码分析和领域知识结合在一起，为理解和快速构建系统提供更好的工具。

保罗·沃特莫（Paul Whatmough）领导马萨诸塞州波士顿ARM研究院的机器学习计算机体系结构研究。他还是哈佛大学工程与应用科学学院的副教授。Whatmough博士在英国兰卡斯特大学

获得了一等荣誉学士学位，在英国布里斯托尔大学获得了杰出硕士学位，在英国伦敦大学学院获得了博士学位。他的研究兴趣包括算法、计算机体系结构和电路。他以前曾领导过多个项目，涉及硬件加速器、机器学习、SoC 架构、数字信号处理（DSP）、制造过程差异容错和电源电压噪声。

古杨·魏（Gu-Yeon Wei）是哈佛大学工程与应用科学学院（SEAS）的电子工程与计算机科学系 Gordon McKay 教授。他分别于 1994 年、1997 年和 2001 年在斯坦福大学获得了电气工程学士、硕士和博士学位。他的研究兴趣涉及计算机系统的多个层次：混合信号集成电路、计算机体系结构和高效硬件的设计工具等。他的研究工作重点是确定这些层次的协同机会，以开发各种系统的节能解决方案，从微型扑翼机器人到物联网/边缘设备的机器学习硬件，再到大型服务器的专用加速器。

大卫·布鲁克斯（David Brooks）是哈佛大学工程与应用科学学院的计算机科学系 Haley Family 教授。在加入哈佛大学之前，他是 IBM T. J. Watson 研究中心的研究人员。Brooks 教授在南加州大学获得电气工程学士学位，在普林斯顿大学获得电气工程硕士和博士学位。他的研究兴趣包括针对高性能和嵌入式系统的弹性和高能效的计算机硬件和软件设计。Brooks 教授是美国电气和电子工程师协会（IEEE）的院士，并获得了多项荣誉和奖项，包括 ACM Maurice Wilkes 奖、ISCA 最具影响力论文奖、NSF CAREER 奖、IBM Faculty Partnership 奖和美国国防高级研究计划局（DARPA）青年教授奖。

CONTENTS

目 录

CHAPTER

1

第1章

引　言

机器学习因其在解决众所周知的人工智能难题上所取得的成功已经占据了新闻的头条。从 DeepMind 的 AlpahGo 对阵人类顶尖围棋选手的决定性胜利，到自动驾驶汽车巡航在城市街道的奇迹，这些方法所造成的影响是长远和广泛的。然而，机器学习的数学和计算基础并不是魔法：这些方法在大半个世纪里逐渐发展起来，并且就像其他领域一样，是计算机科学和数学的一部分。

什么是机器学习呢？一种理解认为这是一种对数据进行编程处理的方式。并不像人类专家针对一些问题设计出明确的解决方案，机器学习的方法是隐式的：人类提供一组规则和数据，而计算机则利用这二者自动得到解决方案。这就将研究和工程的负担从识别特定的一次性解决方案，转移到开发能够应用于各种不同问题的间接方法。虽然这种方法本身也伴随许多挑战，但对于没有已知启发式答案的问题，该方法具有解决问题的潜力，并且可以被广泛采用。

本书关注特定类型的机器学习：神经网络。神经网络可以被宽泛地理解为大脑计算的类比。它们包含大量互连的细小元素从而产生复杂的行为。从头构建一个实用的神经网络超出了人类的能力，因此，就像其他机器学习方法一样，需要依赖于间接方法来构建它们。一个神经网络可能会被给予一组图片并被训练来识别物体，或者被给予一组录音并被训练来转录其内容。然而，可能神经网络最有趣的特征就是它们兴起了多久。今天神经网络收获的成果源于过去数十年的播种。因此，为了将当前的事件放入其历史背景中，本书首先回顾一下神经网络的历史。

1.1 神经网络的兴起和衰落

神经网络很早就已经存在了，然而其过去的发展却有些波折。其早

期的工作（例如 McCulloch 和 Pitts[104]）关注建立类似于生物神经元的数学模型。用硬件再造类似大脑行为的尝试最早出现在 20 世纪 50 年代，最具代表性的工作是 Rosenblatt 的感知器（perceptron）[117]。然而对神经网络的研究兴趣却在逐年消退，多年的乐观热情逐渐幻灭，而这种幻灭又被顽强的坚持再次克服。流行观点的潮流如图 1.1 所示，这些观点与影响当今神经网络的主要事件的时间线叠加在一起。Rosenblatt 于 1957 年创造的热门被 Minsky 和 Papert 的开创性书籍《感知器》[106]所摧毁。在该书中，作者驳斥了过度夸大的现状，并强调了感知器自身的技术限制。曾经众所周知的是，单个感知器甚至无法学习到简单类型的函数，例如异或。在那个时期也有一些其他传闻，认为感知器并没有它自身所说的那么重要，这些传闻主要来自人工智能社区，他们认为感知器过度简化了本领域尝试解决的问题的难度。这些事件促使第一个人工智能的冬天到来，在这个时期，对于机器学习（神经网络和更广泛的人工智能）的研究兴趣和经费资助几乎完全消失。

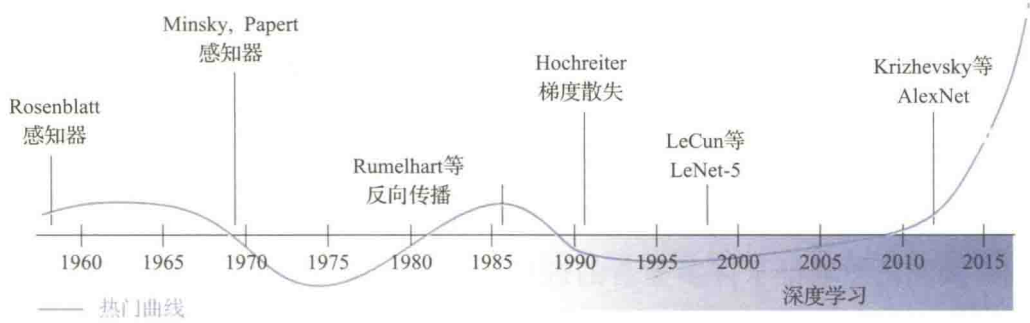


图 1.1 神经网络随时间兴起和衰落的综述。三个主要的巅峰包括：20 世纪 60 年代早期（Rosenblatt）、20 世纪 80 年代中期（Rumelhart）和现在（深度学习）

在 10 年的沉寂后，随着研究者开始认识到过去对神经网络的批评过于严苛，对神经网络的兴趣开始再次升温。新的研究繁荣引入了更大的网络和对其调优的新技术，特别是一种称为并行分布式处理的计算，它可以让大量的神经元同时工作，以实现特定的目标。这 10 年的代表性论

文来自 1986 年由 Rumelhart、Hinton 和 Williams 提出的反向传播方法[119]。虽然也要归功于其他人在更早时期创造该技术（最突出的是 Paul Werbos 在 12 年前提出该技术[140]），但 Rumelhart 等人将反向传播带入主流，并改变了人们对神经网络的态度。反向传播利用简单的微积分方法，从而可以高效地训练任意结构的网络。可能最重要的是，这样可以支持更复杂、层次化的神经网络。反过来，这也扩大了可能被解决的问题系列，并激发了实际应用的兴趣。

尽管有了显著的进展，但过度的热情和炒作再次导致了潜在的问题。实际上，Minsky（他部分煽动和经受过第一个冬天）作为最早几个警醒的人之一曾警告说，如果炒作不消失的话，第二个冬天就可能会到来。为了保持研究经费持续注入，研究人员开始承诺越来越多，当他们无法兑现自己的承诺时，许多资助机构对整个领域的幻想开始破灭。其中值得注意的例子是来自英格兰的 Lighthill 报告[103]，以及 DARPA 取消语言理解研究项目转而资助更加传统的系统。这次衰退伴随着对神经网络复杂性的新认识。Lighthill 特别指出，为了使这些模型在解决实际问题时有用，需要难以置信的大量计算能力，而这在当时是不存在的。

虽然炒作停止了，资助也枯竭了，但研究仍然在后台取得进展。第二个人工智能的冬天从 20 世纪 80 年代晚期延续到 21 世纪头 10 年的中期，其间仍然取得了许多显著的进展。例如，在 20 世纪 90 年代卷积神经网络的开发[89]（见 3.1.1 节），该模型是在更早时期类似的模型上逐步发展起来的（例如，Neocognitron[48]）。然而，又过去了 20 年，对神经网络的广泛兴趣才再次兴起。

1.2 第三波人工智能热潮

在 21 世纪头 10 年的晚期，第二个人工智能冬天开始解冻。虽然神经

网络的算法和理论取得了许多进展，但是使这个时期与众不同的在于神经网络醒来时的环境。作为一个整体，自从20世纪80年代晚期，计算的形势就发生了变化。通过互联网，智能手机普遍连接到社交媒体，产生的数据量激增。同时，计算硬件继续遵循摩尔定律，在整个人工智能冬天呈指数级增长。在20世纪80年代末世界上最强大的计算机差不多等同于2010年的一部智能手机。曾经完全不可行的问题突然看起来变得现实了。

一个良性循环

环境的戏剧性转变开始驱动一个进展与机遇的正循环（图1.2）。正是数据、算法和计算这三个领域的相互作用，直接导致了神经网络的第三次复兴。每个领域自身都很重要，而三者结合在一起的优势就更加深远。

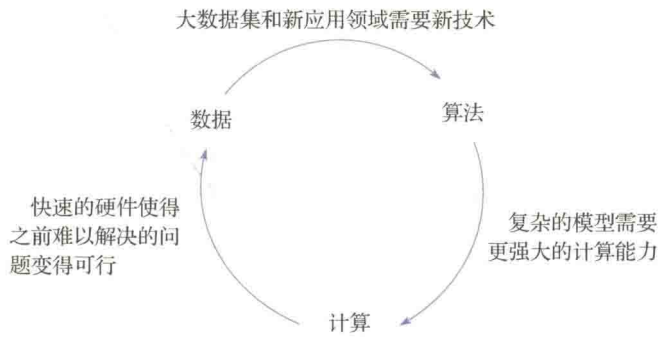


图1.2 良性循环支撑当代深度学习的复兴

这些关键因素形成了一个良性循环。随着更复杂、更大的数据集变得可用，新的神经网络技术被创造出来。这些技术通常涉及更大的模型，并且其机制也要求每个模型参数需要更多的计算量。因此，即便是当今最强大的商业可用的设备，其计算极限也正在受到检验。随着更加强大的硬件被制造出来，模型快速扩展并消耗和使用每一个可用的设备。大数据集、算法训练进步和高性能硬件之间形成一个良性循环的关系。当一个领域取得进步时，就会促进其他两个领域向前发展。