

新时代
技术
新未来

Recommendation
System
and
Deep Learning

本书除了在算法层面讲解推荐系统的实现
还从工程层面详细阐述推荐系统如何搭建

推荐系统 与深度学习

黄昕 赵伟 王本友 吕慧伟 杨敏 —— 编著

清华大学出版社

免费提供配套源代码下载

新时代
技术
新未来

Recommendation
System
and
Deep Learning



推荐系统 与深度学习

黄昕 赵伟 王本友 吕慧伟 何海峰 王斌

清华大学出版社
北京

内 容 简 介

本书的几位作者都曾在大型互联网公司从事与推荐系统相关的实践与研究,通过这本书,把推荐系统工作经验予以总结,以帮助想从事推荐系统工作的读者或推荐系统爱好者。本书的内容设置由浅入深,从传统的推荐算法过渡到近年兴起的深度学习技术。不管是初学者,还是有一定经验的从业人员,相信都能从本书的不同章节中有所收获。

区别于其他推荐算法书籍,本书引入了已被实践证明效果较好的深度学习推荐技术,包括 Word2Vec、Wide & Deep、DeepFM、GAN 等技术应用,并给出了相关的实践代码;除了在算法层面讲解推荐系统的实现,还从工程层面详细阐述推荐系统如何搭建。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

推荐系统与深度学习/黄昕等编著. —北京:清华大学出版社,2019
(新时代·技术新未来)

ISBN 978-7-302-51363-6

I. ①推… II. ①黄… III. ①软件设计 IV. ①TP311.5

中国版本图书馆 CIP 数据核字(2018)第 232192 号

责任编辑:刘洋

封面设计:徐超

责任校对:宋玉莲

责任印制:丛怀宇

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 邮

社总机:010-62770175 邮

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈:010-62772015, zhiliang@tup.tsinghua.edu.cn



印装者:三河市金元印装有限公司

经 销:全国新华书店

开 本:187mm×235mm 印 张:13.75

字 数:247千字

版 次:2019年1月第1版

印 次:2019年1月第1次印刷

定 价:65.00元

产品编号:078535-01

前言

本书的五位作者均曾就职于腾讯，分别在不同的部门从事与推荐系统相关的工作。正是因为“推荐”，我们相识相知。我们不仅在工作中成为伙伴，在工作之余，我们也成了非常好的朋友。在一次好友聊天中，我们萌生写作本书的想法，在之后半年的时间中，我们各有分工，共同完成了本书的写作。可以说，这本书不仅是我们知识的沉淀，也是我们友谊的见证。

推荐算法具有非常多的应用场景和巨大的商业价值。推荐算法种类很多，目前应用最广泛的应该是基于协同过滤的推荐算法。在 2016 年，随着阿尔法围棋（AlphaGo）大放异彩，新的一波深度学习浪潮已至。在图像、音频处理等领域，深度学习技术已成为当之无愧的王者；但在推荐领域，深度学习还处于发展阶段。同时，我们在平时工作学习中，发现市面上并没有关于两者相结合的书籍，只能在国外论文中发现相关的方法与应用。所以，我们决定以比较简单的表达方式，通过总结过往的推荐算法经验，将深度学习相关的应用介绍给更多的读者。

为了适应具有不同知识储备的读者阅读，本书大致可分为四个部分。第 1 至第 3 章为第一部分，主要介绍深度学习的基础知识。第 4 至第 5 章为第二部分，主要介绍了传统的推荐算法及问题。第 6 章为第三部分，进一步介绍深度学习推荐技术。第 7 章为第四部分，介绍了如何在线上实战搭建推荐系统。第三、第四部分是本书的重点，对于从事算法的工作者，可以了解到深度学习技术与推荐算法的结合；对于从事工程的工作者，可以汲取线上搭建推荐系统的经验。

在本书的写作过程中，得到了很多前辈同事的帮助，包括傅鸿城、李深远、刘黎春、赵蕊、邱天宇等领导、同事都给予了很多宝贵意见和支持。没有他们的帮助，我们很难完成本书的写作。

最后还要感谢我们的家人，在写作本书的过程中，我们几位作者占用了大量的家庭时间，感谢他们的照顾和体谅。

目 录

第 1 章 什么是推荐系统	1
1.1 推荐系统的概念	1
1.1.1 推荐系统的基本概念	1
1.1.2 深度学习与推荐系统	4
第 2 章 神经网络	7
2.1 什么是深度学习	7
2.1.1 深度学习的三次兴起	7
2.1.2 深度学习的优势	9
2.2 神经网络基础	11
2.2.1 神经元	11
2.2.2 神经网络	12
2.2.3 反向传播	13
2.2.4 优化算法	14
2.3 卷积网络基础	17
2.3.1 卷积层	17
2.3.2 池化层	19
2.3.3 常见的网络结构	19
2.4 循环网络基础	21
2.4.1 时序反向传播算法	22
2.4.2 长短时记忆网络	24
2.5 生成对抗基础	25
2.5.1 对抗博弈	26
2.5.2 理论推导	27
2.5.3 常见的生成对抗网络	29

第 3 章 TensorFlow 平台	31
3.1 什么是 TensorFlow	31
3.2 TensorFlow 安装指南	33
3.2.1 Windows 环境安装	33
3.2.2 Linux 环境安装	34
3.3 TensorFlow 基础	36
3.3.1 数据流图	36
3.3.2 会话	37
3.3.3 图可视化	37
3.3.4 变量	37
3.3.5 占位符	38
3.3.6 优化器	38
3.3.7 一个简单的例子	38
3.4 其他深度学习平台	39
第 4 章 推荐系统的基础算法	42
4.1 基于内容的推荐算法	42
4.1.1 基于内容的推荐算法基本流程	42
4.1.2 基于内容推荐的特征提取	45
4.2 基于协同的推荐算法	47
4.2.1 基于物品的协同算法	49
4.2.2 基于用户的协同算法	57
4.2.3 基于用户协同和基于物品协同的区别	59
4.2.4 基于矩阵分解的推荐方法	61
4.2.5 基于稀疏自编码的推荐方法	71
4.3 基于社交网络的推荐算法	80
4.3.1 基于用户的推荐在社交网络中的应用	81
4.3.2 node2vec 技术在社交网络推荐中的应用	85
4.4 推荐系统的冷启动问题	94
4.4.1 如何解决推荐系统冷启动问题	94
4.4.2 深度学习技术在物品冷启动上的应用	101

第 5 章 混合推荐系统	119
5.1 什么是混合推荐系统	119
5.1.1 混合推荐系统的意义	120
5.1.2 混合推荐系统的算法分类	122
5.2 推荐系统特征处理方法	125
5.2.1 特征处理方法	126
5.2.2 特征选择方法	134
5.3 常见的预测模型	141
5.3.1 基于逻辑回归的模型	141
5.3.2 基于支持向量机的模型	144
5.3.3 基于梯度提升树的模型	148
5.4 排序学习	150
5.4.1 基于排序的指标来优化	150
5.4.2 L2R 算法的三种情形	152
第 6 章 基于深度学习的推荐模型	156
6.1 基于 DNN 的推荐算法	156
6.2 基于 DeepFM 的推荐算法	163
6.3 基于矩阵分解和图像特征的推荐算法	171
6.4 基于循环网络的推荐算法	174
6.5 基于生成对抗网络的推荐算法	176
6.5.1 IRGAN 的代码实现	179
第 7 章 推荐系统架构设计	183
7.1 推荐系统基本模型	183
7.2 推荐系统常见架构	185
7.2.1 基于离线训练的推荐系统架构设计	185
7.2.2 面向深度学习的推荐系统架构设计	191
7.2.3 基于在线训练的推荐系统架构设计	194
7.2.4 面向内容的推荐系统架构设计	197
7.3 推荐系统常用组件	199
7.3.1 数据上报常用组件	199

7.3.2	离线存储常用组件	200
7.3.3	离线计算常用组件	200
7.3.4	在线存储常用组件	201
7.3.5	模型服务常用组件	201
7.3.6	实时计算常用组件	201
7.4	推荐系统常见问题	201
7.4.1	实时性	201
7.4.2	多样性	202
7.4.3	曝光打击和不良内容过滤	202
7.4.4	评估测试	202
	后记	203

图 目 录

图 1.1	淘宝猜你喜欢栏目	2
图 1.2	百度指数	4
图 1.3	歌曲词嵌入模型空间向量	6
图 2.1	神经网络的三次兴起	8
图 2.2	不同层数的神经网络拟合分界面的能力	10
图 2.3	不同层数的神经网络表示能力	10
图 2.4	神经网络的基本结构	11
图 2.5	感知器算法	12
图 2.6	三层全连接神经网络	13
图 2.7	动量对比	16
图 2.8	卷积运算	18
图 2.9	池化层	19
图 2.10	LeNet 卷积结构	20
图 2.11	Alex-Net 卷积结构	20
图 2.12	RNN	21
图 2.13	LSTM 在 t 时刻的内部结构	24
图 2.14	GAN 网络	25
图 3.1	TensorFlow 安装截图	34
图 3.2	TensorBoard 计算	37
图 4.1	腾讯视频 APP 推荐页面	44
图 4.2	截取自当当网	49
图 4.3	截取自 QQ 音乐 APP	49
图 4.4	用户购买物品记录	50
图 4.5	同时被购买次数矩阵 C	51
图 4.6	相似度计算结果 1	52
图 4.7	相似度计算结果 2	54

图 4.8	相似度计算结果 3	55
图 4.9	截取自当当网	57
图 4.10	物品的倒排索引	57
图 4.11	用户评分矩阵	63
图 4.12	Sigma 值	64
图 4.13	NewData 值	65
图 4.14	Mydata 值	65
图 4.15	自编码神经网络模型	72
图 4.16	稀疏自编码第一个网络	73
图 4.17	稀疏自编码第二个网络	74
图 4.18	稀疏自编码第三个网络	75
图 4.19	将三个网络组合起来	75
图 4.20	社交网络关系图示例	81
图 4.21	融入用户关系和物品关系	82
图 4.22	社交网络关系图示例	86
图 4.23	社交网络关系图示例	86
图 4.24	CBOW 和 Skip-Gram 示例	88
图 4.25	Skip-Gram 网络结构	89
图 4.26	CBOW 网络结构	91
图 4.27	word analogy 示例	93
图 4.28	某网站登录页面	95
图 4.29	QQ 互联开放注册平台 1	96
图 4.30	QQ 互联开放注册平台 2	97
图 4.31	QQ 互联应用管理页面 1	97
图 4.32	QQ 互联应用管理页面 2	97
图 4.33	QQ 互联 QQ 登录功能获取	97
图 4.34	QQ 音乐 APP 中的偏好选择	98
图 4.35	(a) 为每部电影被打分的分布, (b) 为每个用户打分的分布	100
图 4.36	(a) 为每部电影平均分分布, (b) 为每个用户平均分分布	100
图 4.37	基于专家数据的 CF 与基于用户数据 CF 比较	101

图 4.38	音乐频谱示例	102
图 4.39	4 个流派的频谱图示例	103
图 4.40	CNN 音频分类结构	103
图 4.41	CNN+LSTM 组合音频分类模型	104
图 4.42	分类预测结果的混淆矩阵	104
图 4.43	模型倒数第二层 128 维向量降维可视化	104
图 4.44	微软 how-old.net	107
图 4.45	SCUT-FBP 数据集示例图	108
图 4.46	脸部截取后的数据集示例图	108
图 4.47	CNN 层数过多, 误差反而较大	113
图 4.48	残差网络的基本结构	113
图 4.49	残差网络完整结构	114
图 5.1	NetFlix 的实时推荐系统的架构图	120
图 5.2	整体式混合推荐系统	125
图 5.3	并行式混合推荐系统	125
图 5.4	流水线式混合推荐系统	125
图 5.5	MDLP 特征离散化	130
图 5.6	ChiMerge 特征离散化	131
图 5.7	层次化时间按序列特征	133
图 5.8	Learn to rank 的局限	153
图 6.1	Wide & Deep 模型结构	157
图 6.2	推荐系统的召回和排序两个阶段	158
图 6.3	召回模型结构	159
图 6.4	序列信息	160
图 6.5	排序模型结构	161
图 6.6	不同 NN 的效果	162
图 6.7	DeepFM 模型结构 (网络左边为 FM 层, 右边为 DNN 层)	164
图 6.8	FM 一阶部分	165
图 6.9	FM 二阶部分	166
图 6.10	FM/DNN/DeepFM 的比较	171

图 6.11	电影静止帧图片举例	172
图 6.12	Alex-Net 卷积网络	173
图 6.13	左图：时间无关的推荐系统。右图：时间相关的推荐系统	174
图 6.14	基于循环神经网络的推荐系统	175
图 6.15	判别器	177
图 6.16	生成器	178
图 6.17	IRGAN 说明	179
图 7.1	监督学习基本模型	184
图 7.2	基于离线训练的推荐系统架构设计	186
图 7.3	数据上报模块	187
图 7.4	离线训练模块	187
图 7.5	推荐系统中的存储分层	188
图 7.6	在线预测的几个阶段	189
图 7.7	推荐系统通用性设计	190
图 7.8	面向深度学习的推荐系统架构设计	191
图 7.9	利用深度学习进行特征提取	192
图 7.10	参数服务器架构	193
图 7.11	基于在线训练的推荐系统架构设计	195
图 7.12	在线学习之实时特征处理	196
图 7.13	面向内容的推荐系统架构设计	198
图 7.14	用于推荐的内容池	198
图 7.15	Apache Kafka 逻辑架构	200

表 目 录

表 4.1	用户 A 和 B 的评分矩阵	43
表 4.2	电影内容特征二进制表示	45
表 4.3	人脸魅力值打分不同模型的 MAE 比较	112
表 4.4	人脸魅力值打分不同模型的 MAE 比较	117
表 4.5	Keras 预训练好的图像分类模型	118

第 1 章

什么是推荐系统

1.1 推荐系统的概念

1.1.1 推荐系统的基本概念

随着互联网行业发展，我们已进入一个信息爆炸的时代。信息爆炸是互联网赋予当前时代的特征，互联网技术的发展，带给我们最直观的感受是：

1. 各类商品花样繁多；
2. 新闻信息飞速增加；
3. 广告信息铺天盖地；
4. 科技信息迅猛递增；
5. 个人接受力严重“超载”。

以商品的飞速增长为例，美国《连线》主编克里斯·安德森在一篇文章中首次提出了“长尾”（The Long Tail）概念^①：商品销售呈现出长尾形状，冷门商品的需求曲线不会降到零点，而且曲线的尾部比头部长得多，他认为这种丰富性和多样性源自全球化、高效供应链和个性化需求。无独有偶，日本著名营销专家菅谷义博也在《长尾经济学》中提出了这一概念。无论如何，在面对这种多样化和丰富性的时候，都会让人有些手足无措。

另外，随着大数据概念与技术的普及，不管是线上 APP 平台，还是传统的线下平台，都越来越重视数据的收集，数据量呈几何倍数式增长。国际数据公司（IDC）的研究

^① 长尾（The Long Tail）这一概念是由《连线》杂志主编克里斯·安德森（Chris Anderson）在 2004 年 10 月的《长尾》一文中最早提出，用来描述诸如亚马逊和 Netflix 之类网站的商业和经济模式。“长尾”实际上是统计学中幂律（Power Laws）和帕累托分布（Pareto）特征的一个口语化表达。

结果表明, 2008 年全球产生的数据量为 0.49ZB^①, 2009 年的数据量为 0.8ZB, 2010 年增长为 1.2ZB, 2011 年的数量更是高达 1.82ZB, 相当于全球平均每人产生 200GB 以上的数据。IBM 的研究称, 整个人类文明所获得的全部数据中, 有 90% 是过去两年内产生的。而到了 2020 年, 全世界所产生的数据规模将达到今天的 44 倍之多。

信息爆炸与大数据技术的普及, 都促进了个性化推荐技术的快速发展。所谓推荐系统, 简言之就是根据用户的偏好推荐其最有可能感兴趣的内容。以新闻平台为例, 过去主要以新浪新闻这类中心化内容平台为代表; 而现在, 以今日头条为代表的新闻 APP 均在首页根据用户偏好推送不同内容的定制化新闻, 推动了整个行业向个性化推荐转型。在淘宝、京东、亚马逊等电商网站的首页都设有“猜你喜欢”专区, 根据用户最近浏览和购买的行为记录推荐商品。

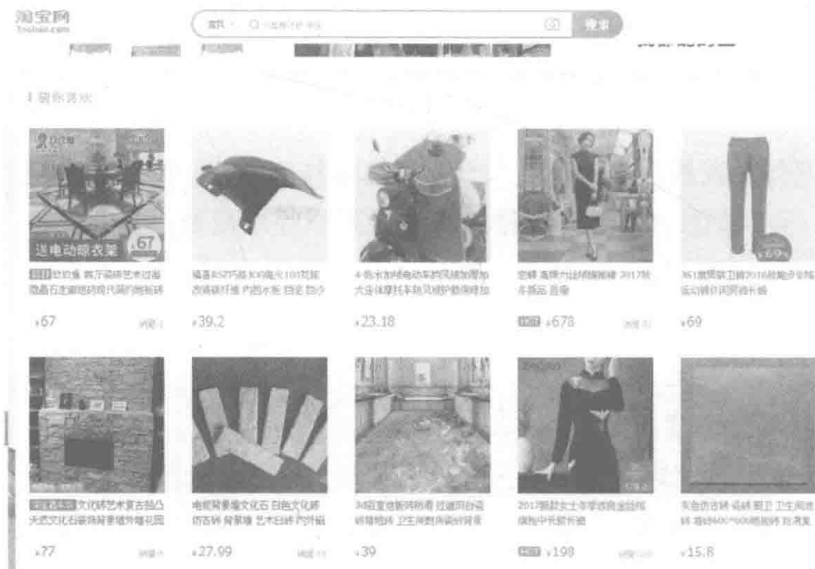


图 1.1 淘宝“猜你喜欢”专区

据数据科学中心 Data Science Central 统计, 对于像亚马逊^②和 Netflix^③这样的主要电子商务平台, 推荐系统可能会承担多达 10% 至 25% 的增量收入。在新兴的短视频领

① ZB, 中文名为泽字节。外文名是 Zettabyte, 计算机信息计量单位, 代表 10^{21} 字节。

② 亚马逊公司 (Amazon, 简称亚马逊; NASDAQ: AMZN), 是美国最大的一家网络电子商务公司, 位于华盛顿州的西雅图。亚马逊是网络上最早开始经营电子商务的公司之一。

③ Netflix (Nasdaq NFLX) 成立于 1997 年, 是一家在线影片租赁提供商, 用户可以通过 PC、TV 及 iPad、iPhone 收看电影、电视节目。Netflix 大奖赛从 2006 年 10 月开始, Netflix 公开了大约 1 亿个 1~5 的匿名影片评级, 比赛要求参赛者预测 Netflix 的客户分别喜欢什么影片, 把预测的效率提高 10% 以上, 推动了推荐技术的发展。

域，以抖音和快手为代表的 APP 以推荐为流量分发的主要手段。在互联网金融领域，各大平台也开始主打针对个人定制化的千人千面投资推荐。毫无疑问，个性化推荐已成为所有新闻、视频、音频、电商、互联网金融等相关平台的标配。面对日益增长的推荐系统需求，推荐系统相关人才的稀缺愈加凸显。

本书第四章是针对推荐系统的初学者，结合算法介绍和实战代码，帮助读者从无到有，从零到一，结构化地掌握推荐系统的基础理论及实践经验。首先，本书会在第四章中介绍最传统的基于内容的推荐算法。这种算法有效利用了推荐内容自身的特点，例如商品的类别标签、新闻的分类标签、音乐的流派标签等，结合用户的历史行为，进行简单有效的推荐。但是这种推荐算法严重依赖物品的内容标签等相关数据，如果完全依靠人工标注，不仅工作量大，准确率也无法保证。所以本书作者结合自身经验，介绍了 TF-IDF 等自动化标签提取方法，以提升基于内容推荐的效率和准确率。推荐系统的一个主流分支是基于协同过滤的推荐算法。该类推荐算法最早由亚马逊提出并应用，目前已成为主流推荐系统的核心技术。基于协同的推荐又可细分为基于记忆的协同过滤和基于模型的协同过滤。其中基于记忆的协同过滤也就是常说的物品协同和用户协同。这两种方法应用最为广泛，但是会过于依赖历史数据，当数据稀疏（例如大部分冷门长尾的商品很少被购买，相关的历史数据就是稀疏的）时，推荐精准度下降严重。另外，随着用户量和商品量的增长，系统性能也会下降。为了解决这一问题，本书会进一步介绍基于模型的协同过滤。最基础的基于模型的协同过滤方法，包括聚类模型和贝叶斯网络等，而目前应用最为广泛的潜在因子推荐方法，被称为奇异值分解（SVD^①）。SVD 可以有效地对用户特征及物品特征进行降维抽象，从而提升系统的效率和准确度；通过进一步考虑用户的隐性行为和时间变化的维度，衍生出了 SVD++、time-SVD++、三维矩阵分解等方法。同时，在互联网化的今天，本书从社交网络的角度去考虑推荐问题，包括基于领域和基于图的社会化推荐方法。

当然推荐系统往往不是由单一算法组成的，混合推荐系统是指将多种推荐技术进行混合，相互弥补缺点，从而可以获得更好的推荐效果。本书的第五章会向读者介绍混合推荐系统中的常用技术。在混合推荐系统或其他机器学习项目中，特征工程是最重要的一个环节，它直接决定了模型的上限。另外，在推荐算法中，除了基于协同算法应用外，还会经常把推荐问题转化为分类问题，例如使用逻辑回归、迭代决策树等分类预测模型进行 CTR 预测。不同于基于物品/用户的协同，分类预测模型更加依赖于特征的处理工

^① 奇异值分解（singular value decomposition）是线性代数中一种重要的矩阵分解，在信号处理、统计学等领域有重要应用。

作及技巧。

在算法理论的基础上，本书将在第七章中结合作者多年工作实践经验，进一步从推荐系统常见架构、常用组件以及常见问题等方面，指导读者如何搭建一个线上高并发可用的推荐系统。

1.1.2 深度学习与推荐系统

深度学习 (Deep Learning) 的概念源于神经网络的研究，是通过探究学习低层特征组合成抽象的高层特征，来解决分类预测问题。深度学习的概念由 Hinton^①等大师于 2006 年提出，主要是使用深度置信网络 (DBN) 进行非监督贪心逐层训练，随后提出多层自动编码器深层结构。另外 Lecun 等大师提出的卷积神经网络 (CNN) 是第一个真正的多层结构学习算法，现在被广泛应用于图像处理领域。2016 年，让国内乃至世界真正认识到深度学习重要性的是阿尔法围棋 (AlphaGo)，它由谷歌 (Google) 旗下 DeepMind 公司戴密斯·哈萨比斯领衔的团队开发。我们从百度搜索指数上也可以看到，深度学习在逐渐赶超机器学习，成为最热门的研究课题。

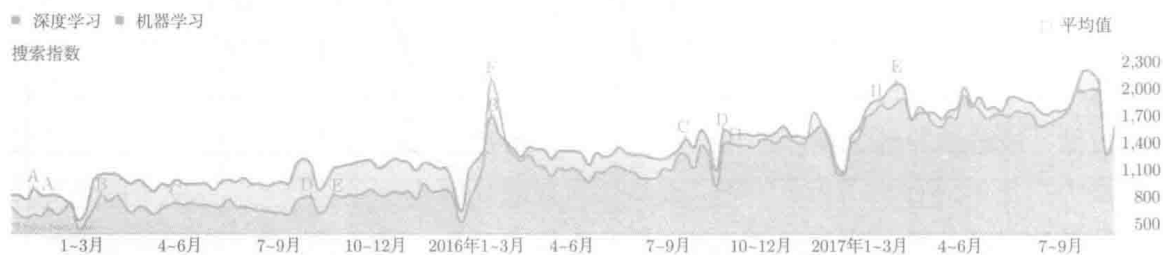


图 1.2 百度搜索指数

严格来说，深度学习只是机器学习的一个研究分支，但之所以被单独挑选出来作为课题研究，是因为深度学习较传统的浅层学习在方法和思维上是一次技术革命。下面具体看一下深度学习的发展历史。

20 世纪 80 年代，用于神经网络的反向传播算法的发明，给机器学习带来了第一次浪潮，它激发大量的学者开始进行机器学习领域的研究。神经网络正是深度学习的

^① Geoffrey Hinton 被尊称为“神经网络之父”，他将神经网络带入到研究与应用的热潮，将“深度学习”从边缘课题变成了谷歌等互联网巨头仰赖的核心技术，并将 Hinton Back Propagation（反向传播）算法应用到神经网络与深度学习，还提出了“Dark Knowledge”概念。曾获得爱丁堡大学人工智能的博士学位，并且为多伦多大学的特聘教授。在 2012 年，Hinton 还获得了加拿大基廉奖 (Killam Prizes，有“加拿大诺贝尔奖”之称的国家最高科学奖)。2013 年，Hinton 加入谷歌并带领一个 AI 团队，目前正进行着谷歌大脑的项目。