Broadview
www.broadview.com.cn

Packt>

在Apache Spark 分布式环境下，提供了80多种简化深度学习的方法

Spark 深度学习指南

[美] Ahmed Sherif　Amrith Ravindra 著
黄友良 译

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Spark
深度学习指南

[美] Ahmed Sherif　Amrith Ravindra 著
黄友良 译

Apache Spark
Deep Learning Cookbook

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

Spark 是专为大规模数据处理而设计的快速通用的计算引擎，经过近几年的飞速发展，现已被广泛应用于各个领域。本书通过通俗易懂的语言和简单明了的操作，系统地讲解了构建 Spark 深度学习系统的方法、流程、标准和规范等相关内容，并提供了相应的示例与解析。

本书适合作为高等院校计算机相关专业的参考资料，也适合大数据技术和机器学习技术的初学者阅读，还适合所有对大数据技术和机器学习技术有所了解并想将该技术应用于本职工作的读者阅读。

版权贸易合同登记号　图字：01-2019-7086

图书在版编目（CIP）数据

Spark 深度学习指南/（美）艾哈迈德 · 谢里夫（Ahmed Sherif），（美）阿姆里斯 · 拉文德拉（Amrith Ravindra）著；黄友良译. —北京：电子工业出版社，2020.1

书名原文：Apache Spark Deep Learning Cookbook

ISBN 978-7-121-37882-9

Ⅰ. ①S… Ⅱ. ①艾… ②阿… ③黄… Ⅲ. ①数据处理软件－指南 Ⅳ. ①TP274-62

中国版本图书馆 CIP 数据核字（2019）第 251443 号

责任编辑：刘恩惠
印　　刷：天津千鹤文化传播有限公司
装　　订：天津千鹤文化传播有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787 × 980　1/16　　印张：23.25　　字数：465 千字
版　　次：2020 年 1 月第 1 版
印　　次：2020 年 1 月第 1 次印刷
定　　价：109.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

译者序

机器学习（Machine Learning）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。它专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。随着计算机软硬件技术的不断发展，机器学习正在悄然改变世界。深度学习（Deep Learning）是机器学习的一个分支和新的研究方向，它能够使计算机通过层次概念来学习经验和理解世界，更接近机器学习实现人工智能的最初目标。深度学习在数据挖掘、机器学习、机器翻译、自然语言处理、个性推荐，以及其他相关领域都取得了很多成果。

在开始深度学习项目之前，选择一个合适的大数据处理框架是非常重要的，选择一个合适的框架能起到事半功倍的作用。Spark 是一个以声明的方式，围绕速度、易用性和复杂分析构建的大数据处理框架，它提供了一个全面、统一的框架，以管理各种有着不同性质（文本数据、图表数据等）的数据集和数据源（批量数据或实时的流数据）。本书用通俗易懂的语言，详细介绍了基于 Spark 的深度学习框架的安装和创建，并通过不同领域的具体案例为读者介绍了 Spark 具体应用方法，希望读者能够更好地理解和应用 Spark。

本书作者 Ahmed Sherif 是一名数据科学家，在美国西北大学获得了预测分析硕士学位。他一直在研究机器学习理论和应用，并使用 Python 和 R 语言进行预测建模。最近，他一直在使用 Azure 开发机器学习和云上深度学习的解决方案。2016 年，他出版了第一本书——《实用商业智能》，他目前是微软公司数据和人工智能领域的技术解决方案专家。本书另外一位作者 Amrith Ravindra 是一名机器学习爱好者，他深入地研究了机器学习理论。Amrith Ravindra 在坦帕市的一次数据科学会议上认识了 Ahmed Sherif，他们决定集思广益，写一

本关于他们最喜欢的机器学习算法的书。他们希望这本书能够帮助大家实现成为一名数据科学家并积极为机器学习做出贡献的最终目标。

非常感谢电子工业出版社计算机出版分社将这样一本好书交给我翻译。我深知自己的知识面和翻译水平有限，难免有疏漏和错误之处，恳请读者批评指正。出版社的编辑帮助我耐心审阅翻译的书稿，并提出了许多中肯的修改意见，非常感谢他们。希望读者能够从书中得到更多的知识，开发更多的机器学习应用来改变世界。

黄友良

2019 年 11 月

序

如果你正在阅读这本书，那么可以肯定的是，你已经意识到**人工智能（AI）**和**机器学习（ML）**对当今的巨大影响，以及深度神经网络带来的不可思议的惊人效率。Matei Zaharia 和他的团队开发的 Spark，不是为了让它成为 Hadoop 的竞争对手，而是作为人工智能和机器学习的广泛应用。正如 Zaharia 所说过的："Spark 只关注如何计算，不关注数据存储在哪里。" Spark 被称为用于大规模数据处理的统一分析引擎，它特别针对弹性、速度、易用性、通用性和可移植性进行了优化，本书将向你详细介绍 Spark，相信你一定会成为 Spark 的狂热爱好者。

作为读者，如果你对 Spark 在深度学习领域的应用感到兴奋，那么本书就可以为你提供帮助。作者首先通过提供简明扼要的文字说明，帮助你设置 Spark 以进行深度学习的开发。在完成初始设置之后，你很自然地就创建了一个神经网络。随后，作者详细说明了卷积神经网络和循环神经网络的难点。在各个行业，人工智能每天都有新的应用。在实际应用方面，作者提供了使用 SparkML 预测消防部门的呼叫，使用 XGBoost 预测房地产价值，使用 LSTM 预测苹果公司的股票价格，以及使用 Keras 创建电影推荐引擎等实际但又进行了适当简化的案例。

人工智能和机器学习分类繁多，令人眼花缭乱的工具集和库的使用令人望而却步。作者非常出色地将不同的库混合在一起，并在读者对本书的学习过程中，不断融入相关但多样化的技术。随着深度学习框架开始融合并向抽象方向发展，探索性数据分析的规模必然会增长。这就是为什么本书并非创造一个众所周知的"一招鲜"（或 YOLO 模型，双关语），而是涵盖了相关或高度相关的技术，如使用长短期记忆单元的生成网络，使用 TF-IDF 的自然语言处理，使用深度卷积网络的人脸识别，使用 Word2Vec 创建和可视化单词向量，并

在 Spark 上使用 TensorFlow 进行图像分类。除简洁和专注的写作风格外，本书的另一个优势是，内容与机器学习、深度学习高度相关。

在学术领域和行业领域，我有一个独特的机会亲自看到 Spark 的发展。作为斯坦福大学的访问学者，我参加了 Spark 的共同创始人之一 Matei Zaharia 的各种会议，并有幸了解了他对 Databricks、大规模算法运作及大数据未来的看法。同时，作为人工智能和深度学习的从业者和首席架构师，我目睹了世界上最大的零售商实验室如何使用 Spark 标准化其业务机器学习模型的部署。随着云技术不断支持新的数据应用程序架构，我认为 Spark 是各个行业的、全新的且不可预料的机器学习实现的关键推动者和加速器。这需要大规模解决现实世界中的业务问题，拥有蓬勃发展的生态系统，并能够为企业级 SLA 提供全天候支持的能力。Spark 以前所未有的速度、弹性和社区适应性满足所有这些标准。Spark 经常用于高级案例分析，如系统遥测的复杂异常值分析。将 Spark 视为一种工具，它可以弥补传统商务智能与现代机器学习服务之间的差距，从而取得有意义的业务成果。随着数据成为核心竞争优势，防患（数据中断）于未然至关重要，同时要考虑季节性、周期性和不断增加的时间相关性。易操作性是行业接纳的关键，而这正是 Spark 与众不同的地方；数据洞察力和行动洞察力也是 Spark 的强项。

你所持有的这本书是 Spark 向人工智能和机器学习广泛应用迈出的又一大步。它为你提供了在实际问题中应用的工具和技术。将它看作探索之旅的开端，成为一个全面的机器学习工程师和传播者，而不仅是一个 Spark 追星族。我们希望这本书能够帮助你利用 Apache Spark 来轻松、高效地处理业务和技术相关的问题。

Adnan Masood 博士
斯坦福大学计算机工程系访问学者
微软最有价值专家&微软人工智能和机器学习首席架构师
斯坦福盖茨计算机科学大楼
2018 年 6 月 12 日

编著者

关于作者

Ahmed Sherif是一名数据科学家，自2005年以来一直从事各种各样的数据研究工作。他从2013年开始使用BI解决方案并慢慢转向数据科学。2016年，他从西北大学获得了预测分析硕士学位，在那里他研究使用Python和R语言进行机器学习和预测建模的科学与应用。最近，他一直使用Azure在云端开发机器学习和深度学习解决方案。2016年，他出版了他的第一本书《实用商业智能》。他目前是微软公司的数据和人工智能技术解决方案专家。

> “首先，我要感谢我的妻子 Ameena 和我的三个可爱的孩子 Safiya、Hamza 和 Layla，感谢他们给我力量和支持来完成这本书。没有他们的爱与支持，我恐怕无法完成本书。我还要感谢我的合著者 Amrith，感谢他给予我写这本书的决心和为本书付出的努力。”

Amrith Ravindra 是一位机器学习爱好者，拥有电气与工业工程学位。在攻读硕士学位期间，他深入地研究了机器学习问题，加深了自己对数据科学的热爱程度。工程专业的研究生课程给他提供了数学背景，使他开始了机器学习领域的职业生涯。他在坦帕市举行的当地数据科学聚会上遇到了 Ahmed Sherif。他们决定合作写一本关于他们最喜欢的机器

学习算法的书。他希望这本书能够帮助他实现成为数据科学家并积极为机器学习做出贡献的最终目标。

> “首先，我要感谢 Ahmed 给我这个机会和他一起工作。对我来说，写这本书比读大学本身更好。接下来，我要感谢我的爸爸、妈妈和姐姐，他们一直给我动力，给我成功的动力。最后，我要感谢我的朋友们，没有他们的指教，我永远不会像这样快速地成长。”

关于审稿人

Michal Malohlava 是 Sparkling Water 的创始人。他是一位极客、开发者，同时也是一位拥有 10 年软件开发经验的 Java、Linux 编程语言爱好者。他于 2012 年在布拉格查理大学获得博士学位，并在普渡大学完成博士后工作。他参与了用于高级大数据数学与计算的 H2O 平台的开发，并将其整合到 Spark 引擎中，以名为 Sparkling Water 的项目发布。

Adnan Masood 博士是人工智能和机器学习研究员、软件架构师和微软数据平台最有价值专家。他目前在 UST Global 公司担任人工智能和机器学习首席架构师，在那里他与斯坦福人工智能实验室和麻省理工学院人工智能实验室合作构建企业解决方案。作为斯坦福大学的访问学者、亚马逊编程语言畅销书 *Functional Programming with F#*的作者，他最近在丹佛市女性技术会议上发表的演讲强调了 STEM（科学、技术、工业、数学）和技术领域多样化的重要性。该演讲被新闻媒体广为传播。

前言

随着深度学习在现代各行业中迅速得到广泛应用，各个机构都在寻找将流行的大数据工具与高效的深度学习库结合起来的方法。这将有助于深度学习模型以更高的效率和更快的速度进行训练。

在本书的帮助下，你将通过学习特定的操作来得到深度学习算法的结果，而不会陷入理论的困境。从为深度学习设置 Apache Spark 到实现各种类型的神经网络，本书解决了大多数常见和不太常见的问题，以便在分布式环境中执行深度学习。除此之外，你还可以访问 Spark 的深度学习代码，这些代码可以用来回答类似问题，也可以在调整后回答稍有不同的问题。你还将学习如何用 Spark 对数据进行流处理和集群处理。一旦掌握了基础知识，你将探索如何使用 TensorFlow 和 Keras 等流行库，如卷积神经网络、循环神经网络和长短期记忆网络，在 Spark 中实现和部署深度学习模型。最后，这是一本旨在教授如何在 Spark 中实际应用模型的指南，所以我们不会深入讨论本书使用的模型背后的理论和数学知识。

在本书的最后，你将拥有在 Apache Spark 上部署高效深度学习模型的专业知识。

本书的读者对象

本书适用于对机器学习和大数据概念有基本了解的人，以及希望通过自上而下而非自下而上的方法来扩展已有知识的人。本书以即学即用的方式进行讲解，任何没有编程经验的人，即使是没有使用过 Python 语言的人，都可以按照提示逐步地轻松实现本书中的算法。本书中的大多数代码都是简单易懂的，每个代码块执行一个特定的功能，或者执行挖掘、转换和将数据拟合到深度学习模型中的操作。

本书旨在通过介绍有趣的项目（如股票价格预测）为读者提供实践经验的同时，让读者对深度学习和机器学习概念有更深入的理解。这可能以提供在线资源链接的方式展现，如已发表的论文、教程和指南，它贯穿本书的每一章。

本书包括哪些内容

第 1 章：为深度学习开发设置 Spark。本章包括在虚拟 Ubuntu 桌面环境下设置 Spark 开发所需的所有内容。

第 2 章：在 Spark 中创建神经网络。本章介绍了从头开始开发神经网络而不使用任何深度学习库（如 TensorFlow 或 Keras）的过程。

第 3 章：卷积神经网络的难点。本章介绍了图像识别中与卷积神经网络相关的一些难点，以及解决问题的方法。

第 4 章：循环神经网络的难点。本章介绍了前馈神经网络和递归神经网络。我们描述了循环神经网络的一些难点，以及如何使用 LSTM 解决它们。

第 5 章：用 Spark 机器学习预测消防部门呼叫。我们将使用 Spark 机器学习开发一个分类模型，用于预测来自旧金山市消防部门的呼叫。

第 6 章：在生成网络中使用 LSTM。本章给出了一种使用小说或大型文本语料库作为输入数据来定义和训练 LSTM 模型的实用方法，同时还使用训练模型生成自己的输出序列。

第 7 章：使用 TF-IDF 进行自然语言处理。本章介绍了分类聊天机器人对话数据升级的步骤。

第 8 章：使用 XGBoost 进行房地产价值预测。本章重点介绍了如何使用金斯县房屋销售数据集来训练一个简单的线性模型，并使用它来预测房价，然后使用一个稍微复杂的模型来做同样的事情并提高预测的准确度。

第 9 章：使用长短期记忆单元预测苹果公司股票的市场价格。本章的重点是使用 Keras 中的 LSTM 创建深度学习模型，以预测苹果公司股票的市场价格。

第 10 章：使用深度卷积网络进行人脸识别。本章利用 10 个不同受试者的面部图像的 MIT-CBCL 数据集来训练和测试深度卷积神经网络模型。

第 11 章：使用 Word2Vec 创建和可视化单词向量。本章重点关注向量在机器学习中的重要性，并指导用户利用谷歌的 Word2Vec 模型训练不同的模型，并可视化小说中产生的单词向量。

第 12 章：使用 Keras 创建电影推荐引擎。本章专注于为使用深度学习库 Keras 的用户

构建电影推荐引擎。

第 13 章：使用 TensorFlow 在 Spark 中进行图像分类。本章专注于利用迁移学习来认识世界知名足球运动员克里斯蒂亚诺·罗纳尔多和里奥·梅西。

如何更好地利用本书

1. 利用书中提供的链接可以更好地理解本书中使用的一些术语。
2. 互联网是当今世界上最大的大学。观看 YouTube、Udemy、edX、Lynda 和 Coursera 等网站提供的有关各种深度学习和机器学习概念的视频。
3. 若仅翻看这本书容易忘记知识点，那么可以在阅读本书时实际执行每一步操作。建议你在浏览每一步操作时打开 Jupyter Notebook，这样就可以在阅读本书时实践每一步操作，同时检查你从每个步骤获得的输出。
4. 本书提供的额外参考资料请访问 http://www.broadview.com.cn/37882 进行下载，如正文中标有参见“链接 1”“链接 2”等字样时，即可从上述网站下载的“参考资料.pdf”文件中进行查询。

下载示例代码文件

你可以从你的账户下载本书的示例代码文件，地址参见链接 1。如果你在其他地方购买了本书，则可以访问链接 2 并通过电子邮件的方式进行注册。

可以按照以下步骤下载示例代码文件。

1. 在链接 3 所示的网址处登录或注册。
2. 选择 **Support** 链接。
3. 单击 **Code Downloads & Errata** 选项。
4. 在 **Search** 框中输入书名，然后按照屏幕上的说明进行操作。

下载文件后，请确保使用最新版本的解压缩软件解压缩文件或文件夹。

- Windows 系统使用 WinRAR/7-Zip。
- Mac 系统使用 Zipeg/iZip/UnRarX。

- Linux 系统使用 7-Zip/PeaZip。

本书的代码包也托管在 GitHub 上；如果代码有更新，则将在现有的 GitHub 存储库上更新：

链接4

我们还提供了丰富的书籍和视频，单击以下链接可进行查看：

链接5

本书的文本约定

本书中使用了许多文本约定。

等宽字体：文本中的代码、数据库表名、文件夹名、文件名、文件扩展名、路径名、虚拟网址、用户输入和 Twitter 句柄均使用等宽字体。下面是一个示例："保存在工作目录中的 `trained` 文件夹下"。

一个代码段如下：

```
print('Total Rows')
df.count()
print('Rows without Null values')
df.dropna().count()
print('Row with Null Values')
df.count()-df.dropna().count()
```

任何命令行的输入或输出都写成了如下的样子：

```
nltk.download("punkt")
nltk.download("stopwords")
```

加粗：表示你在屏幕上看到的新术语、重要单词。例如，菜单或对话框中的选项名称。下面是一个示例："右键单击页面，然后单击 **Save As...**"。

此图标表示警告或重要说明。

此图标表示提示和技巧。

部分

在本书中，你会发现经常出现这样几个标题：准备、实现、说明、扩展、其他，它们的含义如下所示。

准备

本部分将告诉你在操作中需要做什么，并描述如何设置操作所需的软件或初始设置。

实现

本部分包含该操作所需的步骤。

说明

本部分通常包含对“实现”部分中发生的事情的详细说明。

扩展

本部分包含关于操作的附加信息，以便你对操作有更多的了解。

其他

本部分提供了有关操作的其他有用信息链接。

读者服务

微信扫码回复：37882

- 获取博文视点学院 20 元付费内容抵扣券
- 获取本书参考资料中的配套链接
- 获取更多技术专家分享的视频与学习资源
- 加入读者交流群，与更多读者互动

目录