Dipayan Dev

Deep Learning with Hadoop

Build, implement and scale distributed deep learning models for large-scale datasets

Packt>

Deep Learning with Hadoop

This book will teach you how to deploy large-scale dataset in deep neural networks with Hadoop for optimal performance.

Starting with understanding what deep learning is and what the various models associated with deep neural networks are, this book will then show you how to set up the Hadoop environment for deep learning. In this book, you will also learn how to overcome the challenges that you face while implementing distributed deep learning with large-scale unstructured datasets. The book will also show you how you can implement and parallelize the widely used deep learning models such as Deep Belief Networks, Convolutional Neural Networks, Recurrent Neural Networks, Restricted Boltzmann machines and autoencoder using the popular deep learning library Deeplearning4j.

Get in-depth mathematical explanations and visual representations to help you understand the design and implementations of Recurrent Neural Network and Denoising Autoencoders with Deeplearning4j. To give you a more practical perspective, the book will also teach you the implementation of large-scale video processing, image processing and natural language processing on Hadoop.

By the end of this book, you will know how to deploy various deep neural networks in distributed systems using Hadoop.

Things you will learn:

- Explore Deep Learning and various models associated with it
- Understand the challenges of implementing distributed deep learning with Hadoop and how to overcome it
- Implement a Convolutional Neural Network (CNN) with Deeplearning4j
- Delve into the implementation of Restricted Boltzmann machines (RBMs)
- Understand the mathematical explanation for implementing Recurrent Neural Networks (RNNs)
- Understand the design and implementation of Deep Belief Networks (DBN) and Deep Autoencoders using Deeplearning4j
- Get hands on practice of deep learning and its implementation with Hadoop

Packt>
www.packtpub.com

$ **39.99** US
£ **32.99** UK

Prices do not include local sales Tax or VAT where applicable

ISBN 978-1-78712-476-9
53999
9 781787 124769

Deep Learning with Hadoop

Dipayan Dev

Deep Learning with Hadoop

Build, implement and scale distributed deep learning models for large-scale datasets

Dipayan Dev

Packt

BIRMINGHAM - MUMBAI

Deep Learning with Hadoop

Copyright © 2017 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: February 2017

Production reference: 1130217

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.
ISBN 978-1-78712-476-9

www.packtpub.com

试读结束：需要全本请在线购买： www.ertongbook.com

Credits

Authors

Dipayan Dev

Reviewers

Shashwat Shriparv

Wissem EL Khlifi

Commissioning Editor

Amey Varangaonkar

Acquisition Editor

Divya Poojari

Content Development Editor

Sumeet Sawant

Technical Editor

Nilesh Sawakhande

Copy Editor

Safis Editing

Project Coordinator

Shweta H Birwatkar

Proofreader

Safis Editing

Indexer

Mariammal Chettiyar

Graphics

Tania Dutta

Production Coordinator

Melwyn Dsa

About the Author

Dipayan Dev has completed his M.Tech from National Institute of Technology, Silchar with a first class first and is currently working as a software professional in Bengaluru, India. He has extensive knowledge and experience in non-relational database technologies, having primarily worked with large-scale data over the last few years. His core expertise lies in Hadoop Framework. During his postgraduation, Dipayan had built an infinite scalable framework for Hadoop, called Dr. Hadoop, which got published in top-tier SCI-E indexed journal of Springer (`http://link.springer.com/article/10.1631/FITEE.1500015`). Dr. Hadoop has recently been cited by Goo Wikipedia in their Apache Hadoop article. Apart from that, he registers interest in a wide range of distributed system technologies, such as Redis, Apache Spark, Elasticsearch, Hive, Pig, Riak, and other NoSQL databases. Dipayan has also authored various research papers and book chapters, which are published by IEEE and top-tier Springer Journals. To know more about him, you can also visit his LinkedIn profile `https://www.linkedin.com/in/dipayandev`.

About the Reviewers

Shashwat Shriparv has more than 7 years of IT experience. He has worked with various technologies on his career path, such as Hadoop and subprojects, Java, .NET, and so on. He has experience in technologies such as Hadoop, HBase, Hive, Pig, Flume, Sqoop, Mongo, Cassandra, Java, C#, Linux, Scripting, PHP, C++, C, Web technologies, and various real-life use cases in BigData technologies as a developer and administrator. He likes to ride bikes, has interest in photography, and writes blogs when not working.

He has worked with companies such as CDAC, Genilok, HCL, UIDAI(Aadhaar), Pointcross; he is currently working with CenturyLink Cognilytics.

He is the author of *Learning HBase*, Packt Publishing, the reviewer of *Pig Design Pattern* book, Packt Publishing, and the reviewer of *Hadoop Real-World Solution* cookbook, 2nd edition.

I would like to take this opportunity to thank everyone who have somehow made my life better and appreciated me at my best and bared with me and supported me during my bad times.

Wissem El Khlifi is the first Oracle ACE in Spain and an Oracle Certified Professional DBA with over 12 years of IT experience. He earned the Computer Science Engineer degree from FST Tunisia, Masters in Computer Science from the UPC Barcelona, and Masters in Big Data Science from the UPC Barcelona. His area of interest include Cloud Architecture, Big Data Architecture, and Big Data Management & Analysis.

His career has included the roles of: Java analyst / programmer, Oracle Senior DBA, and big data scientist. He currently works as Senior Big Data and Cloud Architect for Schneider Electric / APC. He writes numerous articles on his website `http://www.oracle-class.com` and his twitter handle is `@orawiss`.

www.PacktPub.com

For support files and downloads related to your book, please visit www.PacktPub.com.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

https://www.packtpub.com/mapt

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at `https://www.amazon.com/Deep-Learning-Hadoop-Dipayan-Dev/dp/1787124762`.

If you'd like to join our team of regular reviewers, you can e-mail us at `customerreviews@packtpub.com`. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

试读结束：需要全本请在线购买： www.ertongbook.com

To my mother, Dipti Deb and father, Tarun Kumar Deb.

And also my elder brother, Tapojit Deb.

Table of Contents

Preface 1

Chapter 1: Introduction to Deep Learning 9

- **Getting started with deep learning** 16
  - Deep feed-forward networks 16
  - Various learning algorithms 17
    - Unsupervised learning 19
    - Supervised learning 19
    - Semi-supervised learning 20
- **Deep learning terminologies** 22
- **Deep learning: A revolution in Artificial Intelligence** 25
  - Motivations for deep learning 26
    - The curse of dimensionality 27
    - The vanishing gradient problem 29
    - Distributed representation 31
- **Classification of deep learning networks** 33
  - Deep generative or unsupervised models 35
  - Deep discriminate models 37
- **Summary** 39

Chapter 2: Distributed Deep Learning for Large-Scale Data 41

- **Deep learning for massive amounts of data** 42
- **Challenges of deep learning for big data** 48
  - Challenges of deep learning due to massive volumes of data (first V) 48
  - Challenges of deep learning from a high variety of data (second V) 49
  - Challenges of deep learning from a high velocity of data (third V) 50
  - Challenges of deep learning to maintain the veracity of data (fourth V) 51
- **Distributed deep learning and Hadoop** 51
  - Map-Reduce 53
  - Iterative Map-Reduce 53
  - Yet Another Resource Negotiator (YARN) 54
  - Important characteristics for distributed deep learning design 56
- **Deeplearning4j – an open source distributed framework for deep learning** 58
  - Major features of Deeplearning4j 59
  - Summary of functionalities of Deeplearning4j 60
- **Setting up Deeplearning4j on Hadoop YARN** 60

Getting familiar with Deeplearning4j 60
Integration of Hadoop YARN and Spark for distributed deep learning 66
Rules to configure memory allocation for Spark on Hadoop YARN 66
Summary 71
Chapter 3: Convolutional Neural Network 73
Understanding convolution 74
Background of a CNN 76
Architecture overview 76
Basic layers of CNN 78
Importance of depth in a CNN 79
Convolutional layer 80
Sparse connectivity 80
Improved time complexity 82
Parameter sharing 82
Improved space complexity 82
Equivariant representations 83
Choosing the hyperparameters for Convolutional layers 84
Depth 84
Stride 85
Zero-padding 86
Mathematical formulation of hyperparameters 87
Effect of zero-padding 88
ReLU (Rectified Linear Units) layers 89
Advantages of ReLU over the sigmoid function 90
Pooling layer 91
Where is it useful, and where is it not? 91
Fully connected layer 92
Distributed deep CNN 92
Most popular aggressive deep neural networks and their configurations 93
Training time – major challenges associated with deep neural networks 94
Hadoop for deep CNNs 94
Convolutional layer using Deeplearning4j 97
Loading data 97
Model configuration 98
Training and evaluation 99
Summary 101
Chapter 4: Recurrent Neural Network 103
What makes recurrent networks distinctive from others? 105
Recurrent neural networks(RNNs) 106
Unfolding recurrent computations 106
Advantages of a model unfolded in time 108

Memory of RNNs 109
Architecture 110
Backpropagation through time (BPTT) 112
Error computation 113
Long short-term memory 114
Problem with deep backpropagation with time 114
Long short-term memory 115
Bi-directional RNNs 117
Shortfalls of RNNs 117
Solutions to overcome 118
Distributed deep RNNs 119
RNNs with Deeplearning4j 120
Summary 124
Chapter 5: Restricted Boltzmann Machines 125
Energy-based models 126
Boltzmann machines 128
How Boltzmann machines learn 130
Shortfall 130
Restricted Boltzmann machine 130
The basic architecture 131
How RBMs work 132
Convolutional Restricted Boltzmann machines 135
Stacked Convolutional Restricted Boltzmann machines 137
Deep Belief networks 137
Greedy layer-wise training 138
Distributed Deep Belief network 139
Distributed training of Restricted Boltzmann machines 139
Distributed training of Deep Belief networks 140
Distributed back propagation algorithm 141
Performance evaluation of RBMs and DBNs 141
Drastic improvement in training time 142
Implementation using Deeplearning4j 143
Restricted Boltzmann machines 143
Deep Belief networks 145
Summary 147
Chapter 6: Autoencoders 149
Autoencoder 150
Regularized autoencoders 152
Sparse autoencoders 153

Sparse coding 153
Sparse autoencoders 154
The k-Sparse autoencoder 155
How to select the sparsity level k 156
Effect of sparsity level 156
Deep autoencoders 157
Training of deep autoencoders 158
Implementation of deep autoencoders using Deeplearning4j 160
Denoising autoencoder 163
Architecture of a Denoising autoencoder 163
Stacked denoising autoencoders 164
Implementation of a stacked denoising autoencoder using Deeplearning4j 165
Applications of autoencoders 167
Summary 168
Chapter 7: Miscellaneous Deep Learning Operations using Hadoop 169
Distributed video decoding in Hadoop 170
Large-scale image processing using Hadoop 173
Application of Map-Reduce jobs 174
Natural language processing using Hadoop 175
Web crawler 175
Extraction of keyword and module for natural language processing 177
Estimation of relevant keywords from a page 177
Summary 178
Appendix: References 179
Index 187

试读结束：需要全本请在线购买： www.ertongbook.com