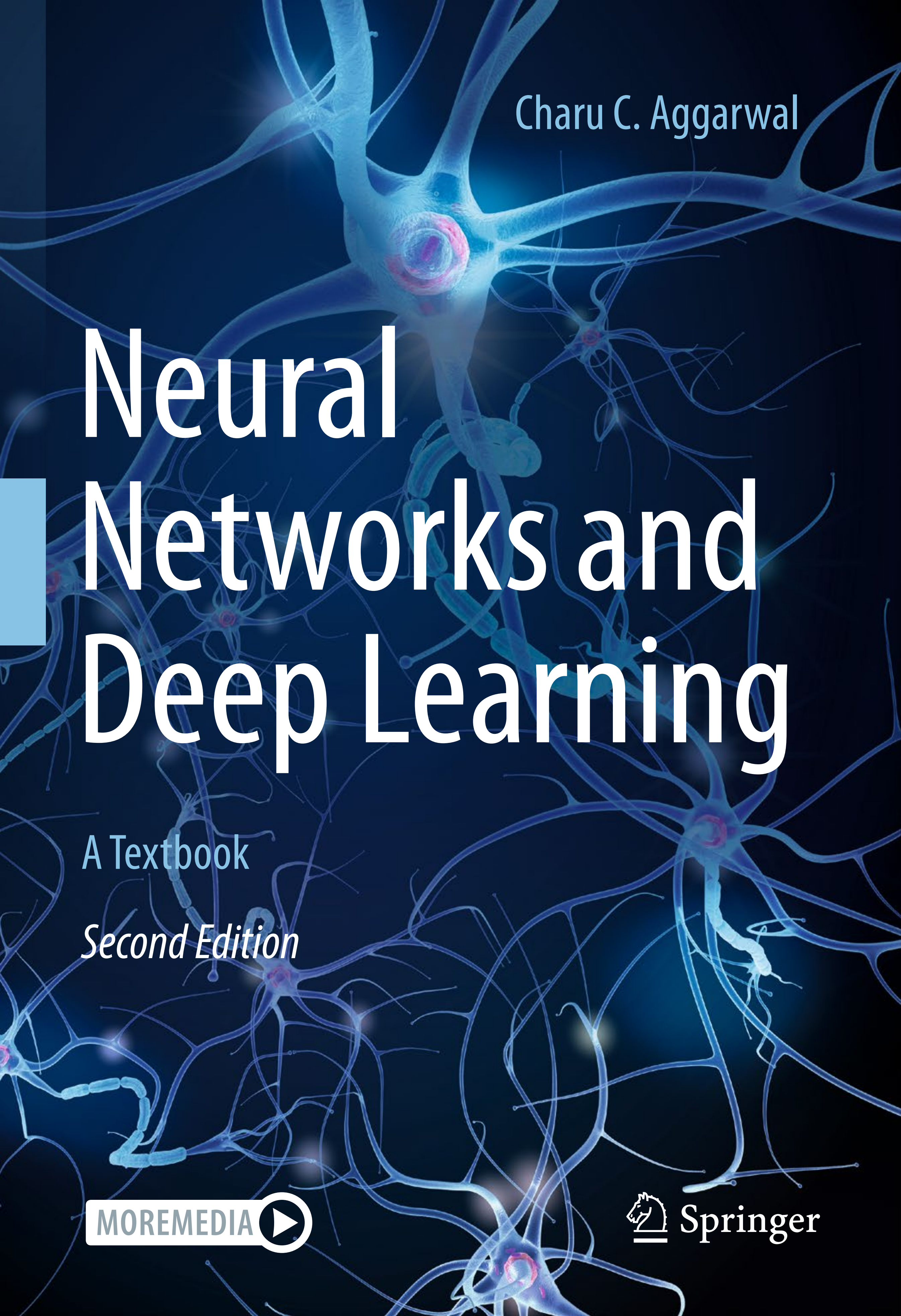

Charu C. Aggarwal

Neural Networks and Deep Learning

A Textbook

Second Edition

MOREMEDIA

Springer

Neural Networks and Deep Learning

Charu C. Aggarwal

Neural Networks and Deep Learning

A Textbook

Second Edition

Springer

Charu C. Aggarwal
IBM T. J. Watson Research Center
International Business Machines
Yorktown Heights, NY, USA

ISBN 978-3-031-29641-3 ISBN 978-3-031-29642-0 (eBook)
https://doi.org/10.1007/978-3-031-29642-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my wife Lata, my daughter Sayani,
and my late parents Dr. Prem Sarup and Mrs. Pushplata Aggarwal.

Preface

> "Any A.I. smart enough to pass a Turing test is smart enough to know to fail it."–***Ian McDonald

Neural networks were developed to simulate the human nervous system for machine learning tasks by treating the computational units in a learning model in a manner similar to human neurons. The grand vision of neural networks is to create artificial intelligence by building machines whose architecture simulates the computations in the human nervous system. Although the biological model of neural networks is an exciting one and evokes comparisons with science fiction, neural networks have a much simpler and mundane mathematical basis than a complex biological system. The neural network abstraction can be viewed as a modular approach of enabling learning algorithms that are based on continuous optimization on a computational graph of mathematical dependencies between the input and output. These ideas are strikingly similar to classical optimization methods in control theory, which historically preceded the development of neural network algorithms.

Neural networks were developed soon after the advent of computers in the fifties and sixties. Rosenblatt's perceptron algorithm was seen as a fundamental cornerstone of neural networks, which caused an initial period of euphoria — it was soon followed by disappointment as the initial successes were somewhat limited. Eventually, at the turn of the century, greater data availability and increasing computational power lead to increased successes of neural networks, and this area was reborn under the new label of "deep learning." Although we are still far from the day that artificial intelligence (AI) is close to human performance, there are specific domains like image recognition, self-driving cars, and game playing, where AI has matched or exceeded human performance. It is also hard to predict what AI might be able to do in the future. For example, few computer vision experts would have thought two decades ago that any automated system could ever perform an intuitive task like categorizing an image more accurately than a human. The large amounts of data available in recent years together with increased computational power have enabled experimentation with more sophisticated and deep neural architectures than was previously possible. The resulting success has changed the broader perception of the potential of deep learning. This book discusses neural networks from this modern perspective. The chapters of the book are organized as follows:

1. *The basics of neural networks:* Chapters 1, 2, and 3 discuss the basics of neural network design and the backpropagation algorithm. Many traditional machine learning models

can be understood as special cases of neural learning. Understanding the relationship between traditional machine learning and neural networks is the first step to understanding the latter. The simulation of various machine learning models with neural networks is provided in Chapter 3. This will give the analyst a feel of how neural networks push the envelope of traditional machine learning algorithms.

2. *Fundamentals of neural networks:* Although Chapters 1, 2, and 3 provide an overview of the training methods for neural networks, a more detailed understanding of the training challenges is provided in Chapters 4 and 5. Chapters 6 and 7 present radial-basis function (RBF) networks and restricted Boltzmann machines.

3. *Advanced topics in neural networks:* A lot of the recent success of deep learning is a result of the specialized architectures for various domains, such as recurrent neural networks and convolutional neural networks. Chapters 8 and 9 discuss recurrent and convolutional neural networks. Graph neural networks are discussed in Chapter 10. Several advanced topics like deep reinforcement learning, attention mechanisms, neural Turing mechanisms, and generative adversarial networks are discussed in Chapters 11 and 12.

We have included some "forgotten" architectures like RBF networks and Kohonen self-organizing maps because of their potential in many applications. An application-centric view is highlighted throughout the book in order to give the reader a feel for the technology.

What Is New in The Second Edition

The second edition has focused on improving the presentations in the first edition and also on adding new material. Significant changes have been made to almost all the chapters to improve presentation and add new material where needed. In some cases, the material in different chapters has been reorganized and reordered in order to improve exposition. The discussion of training challenges with depth has been separated from the backpropagation chapter, so that both topics could be discussed in greater detail. Second-order methods have been explained with greater clarity and examples. Chapter 10 on graph neural networks is entirely new. The discussion on GRUs in Chapter 8 has been greatly enhanced. New convolutional architectures using attention, such as Squeeze-and-Excitation Networks, are discussed in Chapter 9. The chapter on reinforcement learning has been significantly reorganized in order to provide a clearer exposition. The Monte Carlo sampling approach for reinforcement learning is discussed in greater detail in its own dedicated section. Chapter 12 contains expanded discussions on attention mechanisms for graphs and computer vision, transformers, pre-trained language models, and adversarial learning.

Notations

Throughout this book, a vector or a multidimensional data point is annotated with a bar, such as $\overline{X}$ or $\overline{y}$. A vector or multidimensional point may be denoted by either small letters or capital letters, as long as it has a bar. Vector dot products are denoted by centered dots, such as $\overline{X} \cdot \overline{Y}$. A matrix is denoted in capital letters without a bar, such as R. Throughout the book, the $n \times d$ matrix corresponding to the training data set is denoted by D, with n documents and d dimensions. The individual data points in D are therefore d-dimensional row vectors. On the other hand, vectors with one component for each data point are usually n-dimensional column vectors. An example is the n-dimensional column vector $\overline{y}$ of class variables. An observed value y_i is distinguished from a predicted value $\hat{y}_i$ by a circumflex at the top of the variable.

Yorktown Heights, NY, USA

Charu C. Aggarwal

Acknowledgments

Acknowledgements for the First Edition

I would like to thank my family for their love and support during the busy time spent in writing this book. I would also like to thank my manager Nagui Halim for his support during the writing of this book.

Several figures in this book have been provided by the courtesy of various individuals and institutions. The Smithsonian Institution made the image of the Mark I perceptron (cf. Figure 1.5) available at no cost. Saket Sathe provided the outputs in Chapter 8 for the tiny Shakespeare data set, based on code available/described in [243, 604]. Andrew Zisserman provided Figures 9.13 and 9.17 in the section on convolutional visualizations. Another visualization of the feature maps in the convolution network (cf. Figure 9.16) was provided by Matthew Zeiler. NVIDIA provided Figure 11.10 on the convolutional neural network for self-driving cars in Chapter 11, and Sergey Levine provided the image on self-learning robots (cf. Figure 11.9) in the same chapter. Alec Radford provided Figure 12.11, which appears in Chapter 12. Alex Krizhevsky provided Figure 9.9(b) containing *AlexNet*.

This book has benefitted from significant feedback and several collaborations that I have had with numerous colleagues over the years. I would like to thank Quoc Le, Saket Sathe, Karthik Subbian, Jiliang Tang, and Suhang Wang for their feedback on various portions of this book. Shuai Zheng provided feedback on the section on regularized autoencoders in Chapter 5. I received feedback on the sections on autoencoders from Lei Cai and Hao Yuan. Feedback on the chapter on convolutional neural networks was provided by Hongyang Gao, Shuiwang Ji, and Zhengyang Wang. Shuiwang Ji, Lei Cai, Zhengyang Wang and Hao Yuan also reviewed the Chapters 4 and 8, and suggested several edits. They also suggested the ideas of using Figures 9.6 and 9.7 for elucidating the convolution/deconvolution operations.

For their collaborations, I would like to thank Tarek F. Abdelzaher, Jinghui Chen, Jing Gao, Quanquan Gu, Manish Gupta, Jiawei Han, Alexander Hinneburg, Thomas Huang, Nan Li, Huan Liu, Ruoming Jin, Daniel Keim, Arijit Khan, Latifur Khan, Mohammad M. Masud, Jian Pei, Magda Procopiuc, Guojun Qi, Chandan Reddy, Saket Sathe, Jaideep Srivastava, Karthik Subbian, Yizhou Sun, Jiliang Tang, Min-Hsuan Tsai, Haixun Wang, Jianyong Wang, Min Wang, Suhang Wang, Joel Wolf, Xifeng Yan, Mohammed Zaki, ChengXiang Zhai, and Peixiang Zhao. I would also like to thank my advisor James B. Orlin for his guidance during my early years as a researcher.

I would like to thank Lata Aggarwal for helping me with some of the figures created using PowerPoint graphics in this book. My daughter, Sayani, was helpful in incorporating special effects (e.g., image color, contrast, and blurring) in several JPEG images used at various places in this book.

Acknowledgements for the Second Edition

The chapter on graph neural networks benefited from collaborations with Jiliang Tang, Lingfei Wu, and Suhang Wang. Shuiwang Ji, Meng Liu, and Hongyang Gao provided detailed comments on the chapter on graph neural networks as well as other parts of the book. Zhengyang Wang and Shuiwang Ji provided feedback on the new section on transformer networks. Yao Ma and Jiliang Tang provided the permission to use Figure 10.3. Yao Ma also provided detailed comments on the chapter on graph neural networks. Sharmishtha Dutta, a PhD student at RPI, proofread Chapter 10, and also pointed out the sections that needed further clarity. My manager, Horst Samulowitz, provided support during the writing of the second edition of this book.

Contents

Author Biography

Charu C. Aggarwal is a Distinguished Research Staff Member (DRSM) at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his undergraduate degree in Computer Science from the Indian Institute of Technology at Kanpur in 1993 and his Ph.D. from the Massachusetts Institute of Technology in 1996.

He has worked extensively in the field of data mining. He has published more than 400 papers in refereed conferences and journals and authored over 80 patents. He is the author or editor of 20 books, including textbooks on data mining, recommender systems, and outlier analysis. Because of the commercial value of his patents, he has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bio-terrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, and a recipient of two IBM Outstanding Technical Achievement Awards (2009, 2015) for his work on data streams/high-dimensional data. He received the EDBT 2014 Test of Time Award for his work on condensation-based privacy-preserving data mining. He is a recipient of the IEEE ICDM Research Contributions Award (2015) and ACM SIGKDD Innovation Award, which are the two most prestigious awards for influential research contributions in the field of data mining. He is also a recipient of the W. Wallace McDowell Award, which is the highest award given solely by the IEEE Computer Society across the field of Computer Science.

He has served as the general co-chair of the IEEE Big Data Conference (2014) and as the program co-chair of the ACM CIKM Conference (2015), the IEEE ICDM Conference (2015), and the ACM KDD Conference (2016). He served as an associate editor of the IEEE Transactions on Knowledge and Data Engineering from 2004 to 2008. He is an associate editor of the IEEE Transactions on Big Data, an action editor of the Data Mining and Knowledge Discovery Journal, and an associate editor of the Knowledge and Information Systems Journal. He has served or currently serves as the editor-in-chief of the ACM Transactions on Knowledge Discovery from Data as well as the ACM SIGKDD Explorations. He is also an editor-in-chief of ACM Books. He serves on the advisory board of the Lecture Notes on Social Networks, a publication by Springer. He has served as the vice-president of the SIAM Activity Group on Data Mining and is a member of the SIAM industry committee. He received the ACM SIGKDD Service Award for the aforem3entioned contributions to running conferences and journals— this honor is the most prestigious award for services to the field of data mining. He is a fellow of the SIAM, ACM, and the IEEE, for "contributions to knowledge discovery and data mining algorithms."