Lecture Notes in Computer Science 7700

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Grégoire Montavon
Geneviève B. Orr
Klaus-Robert Müller (Eds.)

Neural Networks: Tricks of the Trade

Second Edition

Springer

Volume Editors

Grégoire Montavon
Technische Universität Berlin
Department of Computer Science
Franklinstr. 28/29, 10587 Berlin, Germany
E-mail: gregoire.montavon@tu-berlin.de

Geneviève B. Orr
Willamette University
Department of Computer Science
900 State Street, Salem, OR 97301, USA
E-mail: gorr@willamette.edu

Klaus-Robert Müller
Technische Universität Berlin
Department of Computer Science
Franklinstr. 28/29, 10587 Berlin, Germany
and
Korea University
Department of Brain and Cognitive Engineering
Anam-dong, Seongbuk-gu, Seoul 136-713, Korea
E-mail: klaus-robert.mueller@tu-berlin.de

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-35288-1 e-ISBN 978-3-642-35289-8
DOI 10.1007/978-3-642-35289-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012952591

CR Subject Classification (1998): F.1, I.2.6, I.5.1, C.1.3, F.2, J.3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface to the Second Edition

There have been substantial changes in the field of neural networks since the first edition of this book in 1998. Some of them have been driven by external factors such as the increase of available data and computing power. The Internet made public massive amounts of labeled and unlabeled data. The ever-increasing raw mass of user-generated and sensed data is made easily accessible by databases and Web crawlers. Nowadays, anyone having an Internet connection can parse the 4,000,000+ articles available on Wikipedia and construct a dataset out of them. Anyone can capture a Web TV stream and obtain days of video content to test their learning algorithm.

Another development is the amount of available computing power that has continued to rise at steady rate owing to progress in hardware design and engineering. While the number of cycles per second of processors has thresholded due to physics limitations, the slow-down has been offset by the emergence of processing parallelism, best exemplified by the massively parallel graphics processing units (GPU). Nowadays, everybody can buy a GPU board (usually already available in consumer-grade laptops), install free GPU software, and run computation-intensive simulations at low cost.

These developments have raised the following question: Can we make use of this large computing power to make sense of these increasingly complex datasets? Neural networks are a promising approach, as they have the intrinsic modeling capacity and flexibility to represent the solution. Their intrinsically distributed nature allows one to leverage the massively parallel computing resources.

During the last two decades, the focus of neural network research and the practice of training neural networks underwent important changes. Learning in deep (or "deep learning") has to a certain degree displaced the once more prevalent regularization issues, or more precisely, changed the practice of regularizing neural networks. Use of unlabeled data via unsupervised layer-wise pretraining or deep unsupervised embeddings is now often preferred over traditional regularization schemes such as weight decay or restricted connectivity. This new paradigm has started to spread over a large number of applications such as image recognition, speech recognition, natural language processing, complex systems, neuroscience, and computational physics.

The second edition of the book *reloads* the first edition with more tricks. These tricks arose from 14 years of theory and experimentation (from 1998 to 2012) by some of the world's most prominent neural networks researchers. These tricks can make a substantial difference (in terms of speed, ease of implementation, and accuracy) when it comes to putting algorithms to work on real problems. Tricks may not necessarily have solid theoretical foundations or formal validation. As Yoshua Bengio states in Chap. 19, "the wisdom distilled here should be taken as a guideline, to be tried and challenged, not as a practice set in stone" [1].

The second part of the new edition starts with tricks to faster optimize neural networks and make more efficient use of the potentially infinite stream of data presented to them. Chapter 18 [2] shows that a simple stochastic gradient descent (learning one example at a time) is suited for training most neural networks. Chapter 19 [1] introduces a large number of tricks and recommendations for training feed-forward neural networks and choosing the multiple hyper-parameters.

When the representation built by the neural network is highly sensitive to small parameter changes, for example, in recurrent neural networks, second-order methods based on mini-batches such as those presented in Chap. 20 [9] can be a better choice. The seemingly simple optimization procedures presented in these chapters require their fair share of tricks in order to work optimally. The software *Torch7* presented in Chap. 21 [5] provides a fast and modular implementation of these neural networks.

The novel second part of this volume continues with tricks to incorporate invariance into the model. In the context of image recognition, Chap. 22 [4] shows that translation invariance can be achieved by learning a k-means representation of image patches and spatially pooling the k-means activations. Chapter 23 [3] shows that invariance can be injected directly in the input space in the form of elastic distortions. Unlabeled data are ubiquitous and using them to capture regularities in data is an important component of many learning algorithms. For example, we can learn an unsupervised model of data as a first step, as discussed in Chaps. 24 [7] and 25 [10], and feed the unsupervised representation to a supervised classifier. Chapter 26 [12] shows that similar improvements can be obtained by learning an unsupervised embedding in the deep layers of a neural network, with added flexibility.

The book concludes with the application of neural networks to modeling time series and optimal control systems. Modeling time series can be done using a very simple technique discussed in Chap. 27 [8] that consists of fitting a linear model on top of a "reservoir" that implements a rich set of time series primitives. Chapter 28 [13] offers an alternative to the previous method by directly identifying the underlying dynamical system that generates the time series data. Chapter 29 [6] presents how these system identification techniques can be used to identify a Markov decision process from the observation of a control system (a sequence of states and actions in the reinforcement learning terminology). Chapter 30 [11] concludes by showing how the control system can be dynamically improved by fitting a neural network as the control system explores the space of states and actions.

The book intends to provide a timely snapshot of tricks, theory, and algorithms that are of use. Our hope is that some of the chapters of the new second edition will become our companions when doing experimental work—eventually becoming classics, as some of the papers of the first edition have become. Eventually in some years, there may be an urge to reload again...

September 2012 Grégoire
Klaus

Acknowledgments. This work was supported by the World Class University Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008. The editors also acknowledge partial support by DFG (MU 987/17-1).

References

[1] Bengio, Y.: Practical Recommendations for Gradient-based Training of Deep Architectures. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 437–478. Springer, Heidelberg (2012)

[2] Bottou, L.: Stochastic Gradient Descent Tricks. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 421–436. Springer, Heidelberg (2012)

[3] Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Deep Big Multilayer Perceptrons for Digit Recognition. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 581–598. Springer, Heidelberg (2012)

[4] Coates, A., Ng, A.Y.: Learning Feature Representations with k-means. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 561–580. Springer, Heidelberg (2012)

[5] Collobert, R., Kavukcuoglu, K., Farabet, C.: Implementing Neural Networks Efficiently. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 537–557. Springer, Heidelberg (2012)

[6] Duell, S., Udluft, S., Sterzing, V.: Solving Partially Observable Reinforcement Learning Problems with Recurrent Neural Networks. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 687–707. Springer, Heidelberg (2012)

[7] Hinton, G.E.: A Practical Guide to Training Restricted Boltzmann Machines. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 621–637. Springer, Heidelberg (2012)

[8] Lukoševičius, M.: A Practical Guide to Applying Echo State Networks. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 659–686. Springer, Heidelberg (2012)

[9] Martens, J., Sutskever, I.: Training Deep and Recurrent Networks with Hessian-free Optimization. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 479–535. Springer, Heidelberg (2012)

[10] Montavon, G., Müller, K.-R.: Deep Boltzmann Machines and the Centering Trick. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 621–637. Springer, Heidelberg (2012)

[11] Riedmiller, M.: 10 Steps and Some Tricks to Set Up Neural Reinforcement Controllers. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 735–757. Springer, Heidelberg (2012)

[12] Weston, J., Ratle, F., Collobert, R.: Deep Learning Via Semi-supervised Embedding. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 639–655. Springer, Heidelberg (2012)

[13] Zimmermann, H.-G., Tietz, C., Grothmann, R.: Forecasting with Recurrent Neural Networks: 12 Tricks. In: NN: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 687–707. Springer, Heidelberg (2012)

Table of Contents

Representing and Incorporating Prior Knowledge in Neural Network Training

Tricks for Time Series

Big Learning in Deep Neural Networks

Better Representations: Invariant, Disentangled and Reusable

Identifying Dynamical Systems for Forecasting and Control