

Lecture Notes in Artificial Intelligence

11700

Subseries of Lecture Notes in Computer Science

Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

Founding Editor

Jörg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

More information about this series at <http://www.springer.com/series/1244>


Wojciech Samek · Grégoire Montavon ·
Andrea Vedaldi · Lars Kai Hansen ·
Klaus-Robert Müller (Eds.)


Explainable AI: Interpreting, Explaining and Visualizing Deep Learning




Springer

Editors

Wojciech Samek 
Fraunhofer Heinrich Hertz Institute
Berlin, Germany

Andrea Vedaldi 
University of Oxford
Oxford, UK

Klaus-Robert Müller 
Technische Universität Berlin
Berlin, Germany

Grégoire Montavon
Technische Universität Berlin
Berlin, Germany

Lars Kai Hansen 
Technical University of Denmark
Kgs. Lyngby, Denmark

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-030-28953-9 ISBN 978-3-030-28954-6 (eBook)
<https://doi.org/10.1007/978-3-030-28954-6>

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Explainable AI – Preface

Shortly after the emergence of the first computers, researchers have been interested in developing ‘intelligent’ systems that can make decisions and perform autonomously [6]. Until then, some of these tasks were carried out by humans. Transferring the decision process to an AI system might in principle lead to faster and more consistent decisions, additionally freeing human resources for more creative tasks. AI techniques, such as machine learning, have made tremendous progress over the past decades and many prototypes have been considered for use in areas as diverse as personal assistants, logistics, surveillance systems, high-frequency trading, health care, and scientific research. While some AI systems have already been deployed, what remains a truly limiting factor for a broader adoption of AI technology is the inherent and undoubtable risks that come with giving up human control and oversight to ‘intelligent’ machines [1]. Clearly, for sensitive tasks involving critical infrastructures and affecting human well-being or health, it is crucial to limit the possibility of improper, non-robust, and unsafe decisions and actions [4]. Before deploying an AI system, we see a strong need to validate its behavior, and thus establish guarantees that it will continue to perform as expected when deployed in a real-world environment.

In pursuit of that objective, ways for humans to verify the agreement between the AI decision structure and their own ground-truth knowledge [7] have been explored. Simple models such as shallow decision trees or response curves are readily interpretable, but their predicting capability is limited. More recent deep learning based neural networks provide far superior predictive power, but at the price of behaving as a ‘black-box’ where the underlying reasoning is much more difficult to extract. Explainable AI (XAI) has developed as a subfield of AI, focused on exposing complex AI models to humans in a systematic and interpretable manner.

A number of XAI techniques [2, 3, 5, 8] have been proposed. Some of them have already proven useful by revealing to the user unsuspected flaws or strategies in commonly used ML models. However, many questions remain on whether these explanations are robust, reliable, and sufficiently comprehensive to fully assess the quality of the AI system. A series of workshops have taken place at major machine learning conferences on the topic of interpretable and explainable AI. The present book has emerged from our NIPS 2017 workshop “Interpreting, Explaining and Visualizing Deep Learning ... now what?”. The goal of the workshop was to assess the current state of the research on XAI, and to discuss ways to mature this young field.

Therefore, in essence, this book does not provide final answers to the problem of interpretable AI. It is a snapshot of interpretable AI techniques that have been proposed recently, reflecting the current discourse in this field and providing directions of future development.

Our goal was to organize these contributions into a coherent structure, and to explain how each of them may contribute in the ‘big picture’ of interpretable and

explainable AI. A number of chapters in this book are extensions of the workshop contributions. Other papers are contributions from non-participants that have been added to obtain a more comprehensive coverage of the current research flavors. Each chapter has received at least two peer-reviews and the revised contributions have greatly profited from this process.

The book is organized in six parts:

- Part 1: Towards AI Transparency
- Part 2: Methods for Interpreting AI Systems
- Part 3: Explaining the Decisions of AI Systems
- Part 4: Evaluating Interpretability and Explanations
- Part 5: Applications of Explainable AI
- Part 6: Software for Explainable AI

Although not being able to cover the full breadth of topics, the 22 chapters in this book provide a timely snapshot of algorithms, theory, and applications of interpretable and explainable AI. Many challenges still exist both on the methods and theory side, as well as regarding the way explanations are used in practice. We consider the book an excellent starting point that will hopefully enable future work resolving open challenges of this active field of research.

July 2019

Wojciech Samek
Grégoire Montavon
Andrea Vedaldi
Lars Kai Hansen
Klaus-Robert Müller

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete Problems in AI Safety. arXiv preprint arXiv:1606.06565 (2016)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7), e0130140 (2015)
3. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: IEEE International Conference on Computer Vision (CVPR). pp. 3429–3437 (2017)
4. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R.: Unmasking clever hans predictors and assessing what machines really learn. Nature Communications 10, 1096 (2019)
5. Montavon, G., Samek, W., Müller, K.-R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing 73, 1–15 (2018)
6. Rosenblatt, F.: The perceptron, a perceiving and recognizing automaton (Project Para). Report No. 85-460-1. Cornell Aeronautical Laboratory (1957)

7. Samek, W., Müller, K.-R.: Towards explainable artificial intelligence. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. LNCS, vol. 11700, pp. 5–22. Springer, Cham (2019)
8. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: *European Conference Computer Vision (ECCV)*. pp. 818–833 (2014)

Contents

Part I Towards AI Transparency

1. Towards Explainable Artificial Intelligence. 5
Wojciech Samek and Klaus-Robert Müller
2. Transparency: Motivations and Challenges 23
Adrian Weller
3. Interpretability in Intelligent Systems – A New Concept? 41
Lars Kai Hansen and Laura Rieger

Part II Methods for Interpreting AI Systems

4. Understanding Neural Networks via Feature Visualization: A Survey. 55
Anh Nguyen, Jason Yosinski, and Jeff Clune
5. Interpretable Text-to-Image Synthesis with Hierarchical Semantic Layout Generation. 77
Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee
6. Unsupervised Discrete Representation Learning. 97
Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama
7. Towards Reverse-Engineering Black-Box Neural Networks 121
Seong Joon Oh, Bernt Schiele, and Mario Fritz

Part III Explaining the Decisions of AI Systems

8. Explanations for Attributing Deep Neural Network Predictions 149
Ruth Fong and Andrea Vedaldi
9. Gradient-Based Attribution Methods 169
Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross
10. Layer-Wise Relevance Propagation: An Overview 193
Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller
11. Explaining and Interpreting LSTMs 211
Leila Arras, José Arjona-Medina, Michael Widrich, Grégoire Montavon, Michael Gillhofer, Klaus-Robert Müller, Sepp Hochreiter, and Wojciech Samek

Part IV Evaluating Interpretability and Explanations

12. Comparing the Interpretability of Deep Networks via Network Dissection	243
<i>Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba</i>	
13. Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison	253
<i>Grégoire Montavon</i>	
14. The (Un)reliability of Saliency Methods	267
<i>Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim</i>	

Part V Applications of Explainable AI

15. Visual Scene Understanding for Autonomous Driving Using Semantic Segmentation	285
<i>Markus Hofmarcher, Thomas Unterthiner, José Arjona-Medina, Günter Klambauer, Sepp Hochreiter, and Bernhard Nessler</i>	
16. Understanding Patch-Based Learning of Video Data by Explaining Predictions	297
<i>Christopher J. Anders, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller</i>	
17. Quantum-Chemical Insights from Interpretable Atomistic Neural Networks	311
<i>Kristof T. Schütt, Michael Gastegger, Alexandre Tkatchenko, and Klaus-Robert Müller</i>	
18. Interpretable Deep Learning in Drug Discovery	331
<i>Kristina Preuer, Günter Klambauer, Friedrich Rippmann, Sepp Hochreiter, and Thomas Unterthiner</i>	
19. NeuralHydrology – Interpreting LSTMs in Hydrology	347
<i>Frederik Kratzert, Mathew Herrnegger, Daniel Klotz, Sepp Hochreiter, and Günter Klambauer</i>	
20. Feature Fallacy: Complications with Interpreting Linear Decoding Weights in fMRI	363
<i>Pamela K. Douglas and Ariana Anderson</i>	
21. Current Advances in Neural Decoding	379
<i>Marcel A. J. van Gerven, Katja Seeliger, Umut Güçlü, and Yağmur Güçlütürk</i>	

Part VI Software for Explainable AI

22. Software and Application Patterns for Explanation Methods	399
<i>Maximilian Alber</i>	
Subject Index	435
Author Index	439