Report from Dagstuhl Seminar 19452

Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable

Edited by
Enrico Bertini[1], Peer-Timo Bremer[2], Daniela Oelke[3], and Jayaraman Thiagarajan[4]

1 **NYU – Brooklyn, US,** enrico.bertini@nyu.edu
2 **LLNL – Livermore, US,** bremer5@llnl.gov
3 **Siemens AG – München, DE,** daniela.oelke@siemens.com
4 **LLNL – Livermore, US,** jjayaram@llnl.gov

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19452 "Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable".

Seminar November 3–8, 2019 – http://www.dagstuhl.de/19452
2012 ACM Subject Classification Human-centered computing → Visualization, Computing methodologies → Artificial intelligence, Computing methodologies → Machine learning
Keywords and phrases Visualization, Machine Learning, Interpretability
Digital Object Identifier 10.4230/DagRep.9.11.24

1 Executive Summary

Enrico Bertini (NYU – Brooklyn, US, enrico.bertini@nyu.edu)
Peer-Timo Bremer (LLNL – Livermore, US, bremer5@llnl.gov)
Daniela Oelke (Dep. of Informatics, Siemens AG – München, DE, daniela.oelke@siemens.com)
Jayaraman J. Thiagarajan (LLNL – Livermore, US, jayaram@llnl.gov)

The recent advances in machine learning (ML) have led to unprecedented successes in areas such as computer vision and natural language processing. In the future, these technologies promise to revolutionize everything ranging from science and engineering to social studies and policy making. However, one of the fundamental challenges in making these technologies useful, usable, reliable and trustworthy is that they are all driven by extremely complex models for which it is impossible to derive simple (closed-format) descriptions and explanations. Mapping decisions from a learned model to human perceptions and understanding of that world is very challenging. Consequently, a detailed understanding of the behavior of these AI systems remains elusive, thus making it difficult (and sometimes impossible) to distinguish between actual knowledge and artifacts in the data presented to a model. This fundamental limitation should be addressed in order to support model optimization, understand risks, disseminate decisions and findings, and most importantly to promote trust.

While this grand challenge can be partially addressed by designing novel theoretical techniques to validate and reason about models/data, in practice, they are found to be grossly insufficient due to our inability to translate the requirements from real-world applications into tractable mathematical formulations. For example, concerns about AI systems (e.g., biases) are intimately connected to several human factors such as how information is perceived, cognitive biases, etc. This crucial gap has given rise to the field of *interpretable machine learning*, which at its core is concerned with providing a human user better understanding of the model's logic and behavior. In recent years, the machine learning community, as well as virtually all application areas, have seen a rapid expansion of research efforts in interpretability and related topics. In the process, visualization, or more generally interactive systems, have become a key component of these efforts since they provide one avenue to exploit expert intuition and hypothesis-driven exploration. However, due to the unprecedented speed with which the field is currently progressing, it is difficult for the various communities to maintain a cohesive picture of the state of the art and the open challenges; especially given the extreme diversity of the research areas involved.

The focus of this Dagstuhl Seminar was to convene various stakeholders to jointly discuss needs, characterize open research challenges, and propose a joint research agenda. In particular, three different stakeholders were engaged in this seminar: application experts with unmet needs and practical problems; machine learning researchers who are the main source of theoretical advances; and visualization and HCI experts that can devise intuitive representations and exploration frameworks for practical solutions. Through this seminar, the group of researchers discussed the state of practice, identified crucial gaps and research challenges, and formulated a joint research agenda to guide research in interpretable ML.

Program Overview

The main goal of this Dagstuhl seminar was to discuss the current state and future research directions of interpretable Machine Learning. Because two different scientific communities met, the Machine Learning community and the Visualization community, we started the seminar by discussing and defining important terms and concepts of the field. Afterwards, we split up into working groups to collect answers to the following questions: "*Who needs interpretable machine learning? For what task is it needed? Why is it needed?*". This step was then followed by a series of application lightning talks (please refer to the abstracts below for details).

On the second day, we had two overview talks, one covering the machine learning perspective on interpretability, and the other one the visualization perspective on the topic. Afterwards, we built working groups to collect research challenges from the presented applications and beyond.

The third day was dedicated to clustering the research challenges into priority research directions. The following priority research directions were identified:

- Interpreting Learned Features and Learning Interpretable Features
- Evaluation of Interpretability Methods
- Evaluation and Model Comparison with Interpretable Machine Learning
- Uncertainty
- Visual Encoding and Interactivity
- Interpretability Methods
- Human-Centered Design

On Thursday, the priority research directions were further detailed in working groups. We had two rounds of working groups in which 3, respectively 4, priority research challenges were discussed in parallel by the groups according to the following aspects: problem statement, sub-challenges, example applications, and related priority research directions. Furthermore, all research challenges were mapped into descriptive axes of the problem space and the solution space.

On the last day, we designed an overview diagram that helps to communicate the result to the larger scientific community.

2 Table of Contents

3 Overview of Talks

3.1 Understanding Generative Physics Models with Scientific Priors

Rushil Anirudh (LLNL – Livermore, US)

Joint work of Rushil Anirudh, Jayaraman J. Thiagarajan, Peer-Timo Bremer, Brian K. Spears
Main reference Rushil Anirudh, Jayaraman J. Thiagarajan, Shusen Liu, Peer-Timo Bremer, Brian K. Spears: "Exploring Generative Physics Models with Scientific Priors in Inertial Confinement Fusion", CoRR, Vol. abs/1910.01666, 2019.
URL https://arxiv.org/abs/1910.01666

Modern neural networks are highly effective in modeling complex, multi-modal data and thus have raised significant interested in exploiting these capabilities for scientific applications. In particular, the ability to directly ingest multi-modal, non-scalar data, i.e. images, energy spectra, etc., has proven to be a significant advantage over more traditional statistical approaches. One common challenge for such systems is to properly account for various invariants and constraints to guarantee physically meaningful results, i.e. positive energy, mass conservation, etc. Existing approaches either integrate the physical laws, or rather the corresponding partial differential equations, directly into the training process or add the constraints into the loss function. However, this only works for known constraints that can be explicitly formulated as some differentiable equation in order to be integrated into the neural network training. In practice, not all constraints are known or can be formulated in this manner and explicitly enforcing some constraints while ignoring others is likely to bias the resulting system. Furthermore, constraints are often based on unrealistic assumptions, i.e. physical relationships under some idealized condition, which are not satisfied in the real data. Consequently, strictly enforcing such constraints may produce incorrect results.

In this talk, I explored a few ways in which we can explore, evaluate, and understand the behavior of generative models for scientific datasets. By directly incorporating all known constraints into the loss function, evaluating the constraints post-hoc becomes a self-fulfilling prophecy with the compliance driven largely by the choice of weights in the loss function and a significant potential to over-correct the results. At the same time, most existing metrics are either designed for traditional computer vision problems like Inception scores, FID-scores, or they rely on other global metrics like manifold alignment, which may have little significance in the scientific context. Instead, we propose to use the constraints to evaluate a generative model and show how exploring the data distribution in latent space, i.e. the physics manifold, through the lense of the constraint can provide interesting insights. In particular, we use Inertial Confinement Fusion (ICF) as a testbed problem, with multi-modal data generated from a 1D semi-analytic simulator.

3.2 VIS Perspectives on Interactive and Explainable Machine Learning

Mennatallah El-Assady (Universität Konstanz, DE)

Main reference Thilo Spinner, Udo Schlegel, Hanna Schäfer, Mennatallah El-Assady: "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning", IEEE Trans. Vis. Comput. Graph., Vol. 26(1), pp. 1064–1074, 2020.
URL https://doi.org/10.1109/TVCG.2019.2934629

Interactive and explainable machine learning can be regarded as a process, encompassing thee high-level stages: (1) understanding machine learning models and data; (2) diagnosing model limitations using explainable AI methods; (3) refining and optimizing models interactively.

In my talk, I review the current state-of-the-art of visualization and visual analytics techniques by grouping them into the three stages. In addition, I argue for expanding our approach to explainability through adapting concepts like metaphorical narratives, verbalization, as well as gamification.

I further introduce the explAIner.ai framework for structuring the process of XAI and IML, as well as operationalizing it through a TensoBoard plugin.

Lastly, to derive a robust XAI methodology, I present a survey on XAI strategies and mediums, transferring knowledge and best practices gained from other disciplines to explainable AI.

3.3 Modernizing Supercomputer Monitoring via Artificial Intelligence

Elisabeth Moore (Los Alamos National Laboratory, US)

This talk is an overview of recent advances at Los Alamos National Laboratory regarding the use of machine learning / artificial intelligence to improve management of datacenters and large-scale computing facilities. Three primary projects will be discussed: (1) Anomaly detection in computer-generated text logs, (2) Natural language processing for job outcome prediction, and (3) Effectiveness of telemetry data for predicting node failures.

3.4 Interpretability Applications: Materials Discovery and Recidivism Prediction

Sorelle Friedler (Haverford College, US)

I present two applications where interpretability is important. First, in materials discovery, the goal is to predict the outcome of chemical experiments. Specifically, the problem is framed as a classification problem where the goal is to predict whether a given set of reactants, at specific masses, temperature, and other experimental conditions, will produce a crystal or not. The goal of the chemists involved in the project is to develop and test scientific hypotheses, i.e., to learn as much as possible about science from the machine learning models. Second, in recidivism predictions, the goal is to reduce the number of people detained pre-trial in the U.S. by releasing more defendants determined to be "low risk". The interpretability goals for this task focus on both understanding each step in a model's prediction and understanding potential unfairness (both racism and sexism) in the machine learning models; both are necessary for defense lawyers to best do their job.

3.5 Human in the loop ML

Nathan Hodas (Pacific Northwest National Lab. – Richland, US)

For few-shot learning, the user specifies a small training set (1-5 images or data points) and the system looks for matches. With only a few data points, this allows for ambiguity in the task. In this case, the user needs to “explain” to the computer what the task is (what does it mean to make a good match?). Similarly, the computer needs to explain to the user how it is making decisions, so the user can alter their explanations, in turn.

Sharkzor is used by scientists and other non-data scientists to conduct ML in real-time without any code, so any solution needs to leverage strong human-in-the-loop analytics and minimal friction for interaction. Taken together, HITL explanations and few-shot learning will become increasingly important for non-ML experts to benefit from advanced Machine Learning.

3.6 Application Scenarios for Explainable AI in an Industrial Setting

Daniela Oelke (Siemens AG – München, DE)

In my talk I gave three examples for industrial applications with a need for making machine learning models transparent. In the first example XAI is needed to get a proof that the employed machine learning model takes the right decision in all potential situations of a safety-critical scenario. The second example showcased an application in which the decisions of an anomaly detection system had to be explained. Finally, I presented a use case from the domain of energy management in which the need for calibrated trust and validation was on the focus.

3.7 Explainable AI for Maritime Anomaly Detection and Autonomous Driving.

Maria Riveiro (Univ. of Skövde, SE & Univ. of Jönköning, SE)

Main reference Maria Riveiro: “Evaluation of Normal Model Visualization for Anomaly Detection in Maritime Traffic”, TiiS, Vol. 4(1), pp. 5:1–5:24, 2014.
URL https://doi.org/10.1145/2591511
Main reference Tove Helldin, Göran Falkman, Maria Riveiro, Staffan Davidsson: “Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving”, in Proc. of the Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI ’13, Eindhoven, The Netherlands, October 28-30, 2013, pp. 210–217, ACM, 2013.
URL http://dx.doi.org/10.1145/2516540.2516554

The aim of this talk is to present two application scenarios where visual explanations were provided in order to support users’ decision-making processes.

The first scenario, maritime anomaly detection [1], concerns the analysis of spatio-temporal data to find anomalous behavior in maritime traffic. In this case, machine learning methods were used to create normal behavioral models of different types of vessels. We studied how to present and explain the models created (for understanding and improvement) and the detected anomalies to various stakeholders.

The second scenario, autonomous driving [2], concerns how to present the capability of an autonomous vehicle to drive safely, and the effects that such visual explanations have on driver's performance, acceptance and trust.

These scenarios showcase specific challenges in explainable AI and interpretable machine learning, for instance: (1) constraints related to the limited time to understand the explanations provided, (2) level of detail and content of the explanations given user's goals and tasks, (3) model improvement by domain experts, (4) design for trust calibration and system acceptance, (5) how to represent and visualize normal behavioral models and anomalies and, finally, (6) evaluation metrics and methods for users using explainable AI-systems over time.

References

1 Riveiro, M. (2014). *Evaluation of normal model visualization for anomaly detection in maritime traffic*. ACM Transactions on Interactive Intelligent Systems (TiiS), 4(1), 5.

2 Helldin, T., Falkman, G., Riveiro, M. and Davidsson, S. (2013). *Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving*. Proc. 5th Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications (Automotive'UI 13), Eindhoven, The Netherlands.

3.8 Ada Health GmbH: ExAI in Digital Health

Sarah Schulz (Ada Health – Berlin, DE)

Ada Health GmbH develops a system that is meant to be a health companion. It is created by doctors, scientists, and industry pioneers to bring the future of personalized health to everyone. As digital health is clearly a sector which has to deal with the fact that there might be consequences to decisions made by AI systems, explainability and transparency of machine behaviour and output is inevitable. At Ada Health there are essentially two stages where explanations are needed:

- Ada's knowledge base is manually curated by medical experts. In order to support and accelerate this process, we apply Natural Language Processing methods to extract relevant medical information from unstructured text. To enable the medical expert to refuse or accept a suggestion made by the system they need (visual) explanations to make a decision in a given context.
- Since Ada aims at providing access to medical information to everyone and empowering people to understand their health better, the factors that led to the suggested diagnoses have to be transparent and comprehensible for non-expert users.

3.9 XAI for insurance

Jarke J. van Wijk (TU Eindhoven, NL)

Joint work of Dennis Collaris, Leon Vink, Jarke van Wijk
Main reference Dennis Collaris, Leo M. Vink, Jarke J. van Wijk: "Instance-Level Explanations for Fraud Detection: A Case Study", CoRR, Vol. abs/1806.07129, 2018.
URL http://arxiv.org/abs/1806.07129

I first told a story about transparency, based on my experience with a fine I got for a red light. Fortunately, the evidence showed the light was green, and hence this was fixed easily. Next, I described our experience with fraud detection work for an insurance company. My MSc student Dennis Collaris has worked hard on that, with somewhat puzzling results: different methods give different explanations, and also, practioners did not seem to care [1].

References

1 Dennis Collaris, Leon M. Vink, Jarke J. van Wijk. *Instance-Level Explanations for Fraud Detection: A Case Study*. ICML Workshop on Human Interpretability in Machine Learning, 28-33, 2018.

4 Open problems

4.1 Interpretability for Scientific Machine Learning

Peer-Timo Bremer (LLNL – Livermore, US)

The ability of data driven models to ingest complex, multimodal data types has enabled a new generation of surrogate modeling in many scientific and engineering applications going far beyond previous scalar response functions. However, the black box nature of these models make it challenging to derive actionable insights even from highly accurate and well-tuned models. As a result, interpretability has been recognized as one of the key capabilities to exploit the full power of modern machine learning for scientific discovery.

4.2 Open Questions and Future Directions in Interpretability Research

Sebastian Lapuschkin (Fraunhofer-Institut – Berlin, DE)

Within the last decade, neural network based predictors have demonstrated impressive – and at times super-human – capabilities. This performance is often paid for with an intransparent prediction process, hindering wide-spread adoption of modern machine learning techniques due to scepticism, safety concerns and distrust, or legal demands (see the European Union's extended General Data Protection Regulation act), e.g. in healtcare and industry.

Recognizing the demand for novel and appropriate solutions to the interpretability problem in ML, the explainable artificial intelligence (XAI) community has proposed numerous methods and solutions in recent years. Here, it is essential to note, that each existing approach answers a different aspect of the interpretability question, and consequently no method constitutes a comprehensive solution to the problem as a whole. In addition to that, most approaches are only applicable effectively under specific conditions in terms of data domain, model architecture and model task.

With a plethora of options to choose from (including future developments), and the fact that not every stakeholder is also an XAI domain expert it is important to ask and ultimately answer the following questions (among others):

1. Which methods do the right thing for one's intent, model and application? (I.e., which kind of information does the method provide, and does it synergize well with the model, e.g. wrt. model architecture and task)
2. Can we define a catalogue of (measurable) quality criteria for XAI methods, considering [1] ?
3. How can we generate explanations for non-domain-experts, which includes domain-specific knowledge (to avoid improper interpretation of explanations)?
4. How can we bridge the gap from explanations for individual model predictions to explanations truly characterizing the general model behavior?

4.3 Explainability for affected users. The role of Information Design

Beatrice Gobbo (Politecnico di Milano – Milano, IT)

Purposes of interpretable and explainable machine learning range from debugging models to raise awareness about their social impact, especially when these models are wrong or biased. However, if visual analytics and information visualiSation have been largely used for addressing problems as explainability for the debugging processes, the same means and tools have scarcely been used for raising awareness of machine learning miscalculations among lay users. Taking into account the ethical role of data visualiSation and how much abstraction or approximation could be used when representing inner workings of complex machine learning models, the communication and information designer, together with other professional figures such as computer scientists, can design artifacts able to funnel perception of reliance and doubt of results of these technologies.

Participants

- Rushil Anirudh
LLNL – Livermore, US
- Enrico Bertini
NYU – Brooklyn, US
- Alexander Binder
Singapore University of Technology and Design, SG
- Peer-Timo Bremer
LLNL – Livermore, US
- Mennatallah El-Assady
Universität Konstanz, DE
- Sorelle Friedler
Haverford College, US
- Beatrice Gobbo
Polytechnic University of Milan, IT
- Nikou Guennemann
Siemens AG – München, DE
- Nathan Hodas
Pacific Northwest National Lab. – Richland, US
- Daniel A. Keim
Universität Konstanz, DE
- Been Kim
Google Brain – Mountain View, US
- Gordon Kindlmann
University of Chicago, US
- Sebastian Lapuschkin
Fraunhofer-Institut – Berlin, DE
- Heike Leitte
TU Kaiserslautern, DE
- Yao Ming
HKUST – Kowloon, HK
- Elisabeth Moore
Los Alamos National Laboratory, US
- Daniela Oelke
Siemens AG – München, DE
- Steve Petruzza
University of Utah – Salt Lake City, US
- Maria Riveiro
Univ. of Skövde, SE & Univ. of Jönköning, SE
- Carlos E. Scheidegger
University of Arizona – Tucson, US
- Sarah Schulz
Ada Health – Berlin, DE
- Hendrik Strobelt
MIT-IBM Watson AI Lab – Cambridge, US
- Simone Stumpf
City, University of London, GB
- Jayaraman Thiagarajan
LLNL – Livermore, US
- Jarke J. van Wijk
TU Eindhoven, NL