Preface

Machine learning has been part of Artificial Intelligence since its inception. Only a perfect being need not learn; all others, be they humans or machines, need to learn in order to enhance their capabilities. In the 1980s, learning from examples and modeling human learning strategies have been investigated in concert [490]. The formal statistical basis of many learning methods was put forward later and is still an integral part of machine learning [298]. Neural networks have always been in the toolbox of methods. Integrating all the pre-processing, exploitation of kernel functions, and transformation steps of a machine-learning process into the architecture of a deep neural network increased the performance of this model type considerably [265]. Modern machine learning is challenged by the amount of data and by the demand of real-time inference. This has led recently to an interest in computing architectures and modern processors. For many years, the machine-learning research could take the von Neumann architecture for granted. All algorithms were designed for the classical CPU. Issues of implementation on a particular architecture were ignored. This is no longer possible. The time for independently investigating machine learning and computational architecture is over.

Computing architecture has experienced a similarly rampant development from mainframe or personal computers in the last century to very large compute clusters and ubiquitous computing of embedded systems in the Internet of Things. Cyber-physical systems' sensors produce a huge amount of streaming data that need to be stored and analyzed. Their actuators need to react in real-time. This establishes a close connection with machine learning. Cyber-physical systems and systems in the Internet of Things consist of diverse components, heterogeneous both in hard- and software [470]. Modern multi-core systems, graphic processors, memory technologies, and hardware-software codesign offer opportunities for better implementations of machine-learning models.

Machine learning and embedded systems together now form a field of research that tackles leading edge problems in machine learning, algorithm engineering, and embedded systems. Machine learning today needs to make the resource demands of learning and inference meet the resource constraints of used computer architecture and platforms. A large variety of algorithms for the same learning method and diverse implementations of an algorithm for particular computing architectures optimize learning with respect to resource efficiency while keeping some guarantees of accuracy. To give just one example: the trade-off between a decreased energy consumption and an increased error rate needs to be theoretically shown for training a model and for model inference. Pruning and quantization are ways of reducing the resource requirements by either compressing or approximating the model. In addition to memory and energy consumption, timeliness is an important issue, since many embedded systems are integrated into large products that interact with the physical world. If the results are delivered too late, they may be useless. As a result, real-time guarantees are needed for

https://doi.org/10.1515/9783110785944-203

such systems. To efficiently utilize the available resources, e.g., processing power, memory, and accelerators, with respect to response time, energy consumption, and power dissipation, different scheduling algorithms and resource management strategies need to be developed.

We have dedicated three books to this emerging field of research. They present the results of 12 years of research in 12 projects that were pursued at the TU Dortmund University in the collaborative research center CRC 876 ("Providing Information by Resource Constrained Data Analysis"), funded by the Deutsche Forschungsgemeinschaft (DFG). A collaborative research center is the most selective type of DFG funding. Proposals are submitted in a two-step procedure. The proposals outline a perspective of 12 years in a composition of projects that together shape a research field with a large impact. If this first step is accepted, a detailed proposal for the first phase is submitted and carefully reviewed. After the first phase, its results together with a detailed proposal for the second phase are reviewed and may result in ending the CRC. Otherwise, the second phase starts and at its end, the results and the proposal for the third phase are submitted. At most, three phases are funded. A CRC is a strategic measure of German research funding. The CRC 876 boosted the careers of project leaders. Overall, CRC 876 had 36 project leaders, only 8 of them have been members from the beginning to the end. Hence, career opportunities could be offered to additional colleagues. The CRC 876 with its graduate school boosted the career of Ph.D. students: until 2021, more than 80 dissertations were successfully completed. Uncounted Bachelor and Master theses have been supervised. From this wealth we draw the content of the three books. In addition, guest authors contribute invited chapters.

- The first book establishes the foundations of this new field. It goes through all the steps from the acquisition of data, their summary, and clustering to the different aspects of resource-aware learning.
  Several learning methods are inspected with respect to their resource requirements and how to enhance their scalability on diverse computing architectures: deep neural networks, graph neural networks, tree ensembles, matrix factorization, and probabilistic graphical models.
- The second book is about machine learning for astroparticle and particle physics. Instruments such as the Large Hadron Collider or Cherenkov telescopes or the IceCube gather petabytes of data within which the relevant ones need to be detected, often in real-time, and be stored for further analysis. This builds upon the fundamental issues of the first book and moves into the pipeline of data acquisition, storage, and access, feature extraction, and learning. Here, machine learning is part of the probabilistic rationalism of epistemology. The physical knowledge is encoded in the Monte Carlo simulation and annotates the observations recorded by the instruments. The interpretation of learned models is to enhance physical knowledge. This yields a circle of theory development that is supported by machine learning.

– The third book describes how resource-aware machine-learning methods solve real-world problems in the areas of medicine, industry 4.0, traffic and smart cities, and mission-critical communication.

Each book is self-contained. Together they offer a comprehensive study of machine learning and embedded systems becoming real-time systems, saving energy and offering solutions to other fields. They represent an overview of the state of the art in studying the mutual dependence of machine learning and embedded system design. The presentation of this overview has been made feasible by an early vision of the importance of linking the two domains. We are enthusiastic about the fact that the vision underlying the creation of CRC 876 has become a main line of research worldwide. An early start has allowed us to study the links intensively. Now we would like to entrust to novices and masters alike what we have learned along the long journey of CRC 876, hoping that they might be inspired to work implementations of machine learning and embedded systems.

Enjoy!
Katharina Morik
Peter Marwedel