DEEM 2019: Workshop on Data Management for End-to-End Machine Learning

Sebastian Schelter
New York University
sebastian.schelter@nyu.edu

Neoklis Polyzotis
Google
npolyzotis@google.com

Manasi Vartak
MIT
mvartak@csail.mit.edu

Stefan Seufert
Amazon Research
seufert@amazon.com

ABSTRACT

The DEEM workshop brings together researchers and practitioners at the intersection of applied machine learning, data management and systems research, with the goal to discuss the arising data management issues in machine learning application scenarios.

ACM Reference Format:
Sebastian Schelter, Neoklis Polyzotis, Manasi Vartak, and Stefan Seufert. 2019. DEEM 2019: Workshop on Data Management for End-to-End Machine Learning. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3299869.3323598

1 INTRODUCTION

Applying Machine Learning (ML) in real-world scenarios is a challenging task. In recent years, the main focus of the database community has been on creating systems and abstractions for the efficient training of ML models on large datasets. However, model training is only one of many steps in an end-to-end ML application, and a number of orthogonal data management problems arise from the large-scale use of ML, which require the attention of the data management community [3, 4].

For example, data preprocessing and feature extraction workloads result in complex pipelines that often require the simultaneous execution of relational and linear algebraic operations. Next, the class of the ML model to use needs to be chosen, for that often a set of popular approaches such as linear models, decision trees and deep neural networks have to be tried out on the problem at hand. The prediction quality of such ML models heavily depends on the choice of features and hyperparameters, which are typically selected in a costly offline evaluation process, that poses huge opportunities for parallelization and optimization. Afterwards, the resulting models must be deployed and integrated into existing business workflows in a way that enables fast and efficient predictions, while still allowing for the lifecycle of models (that become stale over time) to be managed. Managing this lifecycle requires careful bookkeeping of metadata and lineage [5, 8] ("Which data was used to train this model?", "Which models are affected by changes in this feature?") and involves methods for continuous analysis, validation, and monitoring of data and models in production [1, 2, 6, 7]. As a further complication, the resulting systems need to take the target audience of ML applications into account; this audience is very heterogeneous, ranging from analysts without programming skills that possibly prefer an easy-to-use cloud-based solution on the one hand, to teams of data processing experts and statisticians developing and deploying custom-tailored algorithms on the other hand.

2 OBJECTIVE AND TOPICS OF INTEREST

The workshop solicits regular research papers describing preliminary and ongoing research results. In addition, the workshop encourages the submission of industrial experience reports of end-to-end machine learning deployments. Areas of particular interest for the workshop include:

- Data Management in ML Applications
- Definition and Optimization of Complex ML Pipelines
- Systems for Managing the Lifecycle of ML Models
- Systems for Efficient Hyperparameter Search
- ML Services in the Cloud
- Modeling and Provenance of ML Experimentation data
- Integration of ML and Dataflow Systems
- Sourcing, Labeling, Integrating, and Cleaning Data for ML
- Benchmarking of ML Applications

SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5643-5/19/06.
https://doi.org/10.1145/3299869.3323598

- Interpretability and Reproducibility in ML Applications
- Responsible Data Management

Review Process. The workshop accepts regular research papers and industrial papers. Submissions can be short papers (4 pages) or long papers (up to 10 pages). Each paper is reviewed by three PC members in a single-blind manner. Researchers submitting papers are required to mark conflicts of interest with the program committee, and reviewing is handled accordingly. The workshop does not accept submissions from direct colleagues of the chairs.

3 ORGANISATION AND PROGRAM

Organizers. The workshop is chaired by the following members of the research community:

Sebastian Schelter is a Moore-Sloan Data Science Fellow at the Center for Data Science of New York University. His research focuses on the intersection of systems, data management and machine learning. Sebastian spent the last three years as a Senior Applied Scientist with Amazon Core AI in Berlin. He received his Ph.D. from TU Berlin, advised by Volker Markl. During his studies, he has been interning at IBM Research Almaden and Twitter in California. Furthermore, he is engaged in Open Source as a member of the Apache Software Foundation, where he has been involved in the Mahout, Giraph, Flink and MXNet projects, and currently mentors the Apache TVM deep learning project during its incubation phase.

Neoklis Polyzotis is a researcher at Google Research, where he currently leads the data management projects in Google's TensorFlow Extended (TFX) platform for production-grade machine learning. His interests include data management for machine learning, enterprise data search, and interactive data exploration. Before joining Google, he was a professor at UC Santa Cruz. He has received a Ph.D. in Computer Science from the University of Wisconsin at Madison and a diploma in engineering from the National Tech. University of Athens, Greece.

Manasi Vartak recently received her Ph.D. from the MIT Database Group under the supervision of Sam Madden. She conducted research on novel systems to support machine learning and data visualization. Most recently, she has been working on ModelDB, a system to manage ML models, and SeeDB, a system to provide visualization recommendations. In the past, she worked/interned at Microsoft, Google, Facebook, and Twitter, and has been a recipient of the Facebook Ph.D. Fellowship (2016) and the Google Anita Borg Scholarship (2013).

Stephan Seufert is a Software Development Engineer with Amazon Search in Berlin. He obtained his Ph.D. in the area of large-scale graph processing from Saarland University in 2015, advised by Gerhard Weikum. Previously, he held positions as a postdoctoral researcher in the Databases and Information Systems group at Max Planck Institute for Informatics in Germany, and as a Senior Software Engineer at Trifacta, a startup company focused on large-scale data cleaning.

Steering Committee. The steering committee comprises of Markus Weimer from Microsoft AI, Volker Markl from Technische Universität Berlin, and Juliana Freire from New York University.

Invited Speakers. The academic Keynote on "Software Engineering 2.0 for Software 2.0: Towards Data Management for Statistical Generalization" will be given by Ce Zhang from ETH Zürich. The industrial keynote on "Distributed Training of Deep Learning Models for Recommendation Systems" will be presented by Leonidas Galanis from Facebook AI. Additionally, the workshop will feature two invited talks: Julia Stoyanovich (New York University) on "TransFAT: Translating Fairness, Accountability and Transparency into Data Science Practice", and Zachary Lipton (Carnegie Mellon University) on "The Social Impacts of Algorithmic Decision Making: Foundations, Solutions, and Roadblocks".

Acknowledgements. This work was supported by the Moore-Sloan Data Science Environment at New York University.

REFERENCES

[1] Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, et al. 2017. Tfx: A tensorflow-based production-scale machine learning platform. *KDD* (2017), 1387–1395.

[2] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Whang, and Martin Zinkevich. 2019. Data Validation for Machine Learning. *SysML* (2019).

[3] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2017. Data management challenges in production machine learning. *SIGMOD* (2017), 1723–1726.

[4] Sebastian Schelter, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert, and Gyuri Szarvas. 2018. On Challenges in Machine Learning Model Management. *Data Engineering* (2018), 5.

[5] Sebastian Schelter, Joos-Hendrik Boese, Johannes Kirschnick, Thoralf Klein, and Stephan Seufert. 2017. Automatically tracking metadata and provenance of machine learning experiments. *Machine Learning Systems workshop at NeurIPS* (2017).

[6] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating large-scale data quality verification. *PVLDB* 11, 12 (2018), 1781–1794.

[7] Manasi Vartak, Joana M F da Trindade, Samuel Madden, and Matei Zaharia. 2018. Mistique: A system to store and query model intermediates for model diagnosis. *SIGMOD* (2018), 1285–1300.

[8] Manasi Vartak and Samuel Madden. 2018. MODELDB: Opportunities and Challenges in Managing Machine Learning Models. *Data Engineering* (2018), 16.