From Data Science to Production ML

Introducing USENIX OpML

NISHA TALAGALA, BHARATH RAMSUNDAR, AND SWAMINATHAN SUNDARARAMAN

Nisha Talagala is co-founder, CTO/VP of Engineering at ParallelM, a startup focused on production machine learning. Nisha has more than 15 years of expertise in software, distributed systems, machine learning, persistent memory, and flash. Nisha earned her PhD at UC Berkeley on distributed systems research. Nisha holds 63 patents in distributed systems, algorithms, networking, memory architecture, and performance. Nisha is a frequent speaker at both industry and academic conferences and serves on multiple technical conference steering and program committees. She is the Program co-chair for OpML '19. nisha@gprof.com

Bharath Ramsundar did his PhD in computer science at Stanford University where he studied the application of deep-learning to problems in drug discovery. While there, he created the deepchem.io open-source drug discovery project and the moleculenet.ai benchmark suite. Bharath is the co-author of *TensorFlow for Deep Learning: From Linear Regression to Reinforcement Learning* and the forthcoming *Deep Learning for the Life Sciences* with O'Reilly Media. As a co-founder of Computable, Bharath is focused on designing the decentralized protocols that will unlock data and AI to create the next stage of the Internet. bharath.ramsundar@gmail.com

Swaminathan (Swami) Sundararaman is the Lead Architect of ParallelM, an early stage startup focused on production machine learning and deep learning. Swami was previously at Fusion-io, Inc. and Sandisk Corp. He holds a PhD from the University of Wisconsin-Madison. swaminathan.sundararaman@gmail.com

In this article we explain the challenges with deploying ML/DL models in production and how USENIX OpML can help bring participants for different disciplines to address the herculean task of safely managing the model life cycle in production.

Machine learning (ML) and its variants such as deep learning (DL) and reinforcement learning are starting to impact every commercial industry. The 2019 USENIX Conference on Operational Machine Learning (OpML '19), dedicated to operational machine learning and its variants, will focus on the full life cycle of deploying and managing ML into production. The goal of the conference is to help develop robust practices for scaling the management of models (i.e., artifact of learning from big data) throughout their life cycle. Through such practices, we can help organizations transition from manually hand-holding to automated management of ML models in production (i.e., ML version of the move in server operations from "pets to cattle" [9]).

Having engaged with hundreds of data scientists over the past few years, it was clear to us that while generating machine-learning models has become easier, moving them into production still remains challenging. It made us carefully think about the question, what is making machine learning more accessible on the one hand, but challenging for broad deployment on the other?

ML technologies have been around for many decades, with intermittent spikes of activity and interest. In the last few years, however, ML and DL technologies have been proven to work effectively in real world use cases in many domains. This shift is driven by several factors:

- **The Data**: Devices from sensors to robots are generating increasing amounts of rich data (from simple value time series to images, sound, and video). While the data itself is valuable, its ultimate benefit to a business's bottom line comes from the analytics that extract the insights hidden within. While simple data sets (such as streams of individual values) can be analyzed via database queries or complex event-processing techniques, the increasing richness of data (multiple correlated mixed type streams, images, sound, video) requires more complex ML and DL approaches. The increased volumes of data also enable ML/DL algorithms to achieve peak efficiency.
- **The Compute**: The ubiquity of high performance commodity computing, driven by both massive core count increases in individual CPUs and low-cost cloud computing services, have made it possible to match data growth with similarly scalable ML and DL capabilities. Hardware innovations such as GPUs, custom FPGAs, and instruction-set support in modern CPUs have further improved ML algorithm performance, making it practical to train using massive data sets [1].
- **The Algorithms**: The availability of open source algorithms for ML and DL via libraries for analytic engines like Spark, TensorFlow, Caffe, NumPy, scikit-learn [2], just to name a few, now offers a massive range of algorithmic techniques for the data scientist sandbox. With open source, even the most state-of-the-art algorithms in research are frequently publicly available to test, tune, and use, nearly as soon as they are invented.

From Data Science to Production ML: Introducing USENIX OpML

These trends addressed the first issues impeding real-world ML (the data, the compute, and quality algorithmic implementations). The next problem was finding a data scientist to match the specific business problem and data set to a suitable algorithm. A lot has been written about the shortage of data scientists [3]. This issue, while real, has been actively addressed in the last several years with online data science courses, specialty programs in universities for data science, and tools that simplify model creation (the democratization of data science) [5]. The latest approach to mitigating this problem, AutoML [4], promises to automate the process of model creation and selection, making it even easier to improve the productivity of a single data scientist.

These trends have also helped generate lots of models. However, to be useful for any application, the model has to be deployed in production with its outputs (recommendations, classifications, etc.) connected to the application that needs it. Deploying, managing, and optimizing ML/DL in production incurs additional challenges:

- **Real-World Dynamism**: Depending on use case, incoming data feeds can change dramatically, possibly beyond what was evaluated in the data scientist sandbox. This in turn affects production ML behavior in ways that are hard to predict or detect via standard production means.
- **Expertise Mismatch**: On one side, IT operations administrators are experts in deployment and management of software and services in production. On the other side, data scientists are experts in the algorithms and associated mathematics. Operating ML/DL in production requires the combined skills of both groups.
- **Non-Intuitive Complexity**: In contrast to other intuitive analytics like rule-based, relational database or pattern matching key-value-based systems (where the output can be predicted from the input values), the core of ML/DL algorithms are mathematical functions (i.e., models) whose data-dependent behavior is not intuitive to most humans.
- **Reproducibility and Diagnostics Challenges**: Since ML/DL algorithms can be probabilistic in nature, there is no consistently "correct" result. For example, even for the same data input, many different outputs are possible depending on what recent training occurred and other factors (such as parameters used to train a model).
- **Inherent Heterogeneity**: Many classes of ML algorithms exist (e.g., machine learning, deep learning, reinforcement learning), and specialized analytic engines (Spark, TensorFlow, PyTorch, containers to train/serve models via Kubernetes) have emerged, each excelling at some subset [2]. Practical ML solutions frequently combine different algorithmic techniques, requiring the production deployment to leverage multiple engines. This makes the deployment process even more fragile than the current data ingestion and processing pipelines. This is uncommon in other application spaces. In databases, for example, standardizing on a single type of DB for a workflow can be a useful production norm.

The term *Cambrian explosion* has already been used in several contexts to describe the growth of AI [6, 7]. Within this trend, what we are seeing now is the explosion of models in the data scientist sandbox, models that cannot be practically used until they are able to deliver on their promise in production. As the number of data scientists increases, as democratization and AutoML tools improve data science productivity, and as compute power grows making it easier to test new algorithms in sandbox, more and more models will be developed, each one awaiting the move into production use.

To help meet this challenge and support the growing community of ML researchers and engineers, data scientists, IT and DevOps engineers who are working to manage ML in production, several of us in industry have worked with USENIX to launch the first conference dedicated to Operational Machine Learning (OpML).

The goal of this conference is to bring the research and industry technical communities together to develop and bring to practice impactful research advances and cutting edge solutions to this problem. Unlike existing conferences and workshops, OpML will focus on "the final stage of deploying and managing ML into production and the subsequent continuous ML/DL lifecycle in production." This covers deployment, automation, orchestration, monitoring, diagnostics, compliance, governance, and the challenges of safely operating and optimizing production systems running ML/DL/Advanced algorithms on live data.

OpML will also provide several benefits for industry and academic participants (please see CFP for details in [8]). Submissions were due on February 15, 2019.

We invite you to participate in the inaugural OpML conference that will be held on May 20, 2019, in Santa Clara, CA, USA.

References

[1] "Nvidia Morphs from Graphics and Gaming to AI and Deep Learning," ZDNet, September 8, 2017: https://www.zdnet.com/article/nvidia-morphs-from-graphics-and-gaming-to-ai-and-deep-learning/.

[2] M. Heller, "Review: The Best Frameworks for Machine Learning and Deep Learning," InfoWorld, February 1, 2017: https://www.infoworld.com/article/3163525/analytics/review-the-best-frameworks-for-machine-learning-and-deep-learning.html.

[3] V. Zhang and C. Neimeth, "Three Reasons Why Data Scientist Remains the Top Job in America," InfoWorld, April 14, 2017: https://www.infoworld.com/article/3190008/big-data/3-reasons-why-data-scientist-remains-the-top-job-in-america.html.

[4] AutoML: http://www.ml4aad.org/automl/.

[5] M. Dillon, "The Democratization of Data Science and the Emergence of Citizen Data Scientists," *Daily Californian,* May 26, 2017: http://www.dailycal.org/2017/05/26/democratization-data-science-emergence-citizen-scientists/.

[6] G. Leopold, "Nvidia CEO Predicts AI 'Cambrian Explosion,'" HPC Wire, May 25, 2017: https://www.hpcwire.com/2017/05/25/nvidia-ceo-predicts-ai-cambrian-explosion/.

[7] S. Condon, "Google's Fei-Fei Li: Vision Is AI's 'Killer App,'" ZDNet, May 19, 2017: https://www.zdnet.com/article/googles-fei-fei-li-vision-is-ais-killer-app/.

[8] USENIX OpML Call for Participation: https://www.usenix.org/sites/default/files/opml19_cfp_121319.pdf.

[9] R. Bias, "The History of Pets vs Cattle and How to Use the Analogy Properly," September 29, 2016: http://cloudscaling.com/blog/cloud-computing/the-history-of-pets-vs-cattle/.