OpML '19: 2019 USENIX Conference on Operational Machine Learning
May 20, 2019
Santa Clara, CA, USA

Production Experiences and Learnings

Opportunities and Challenges Of Machine Learning Accelerators In Production 1
Rajagopal Ananthanarayanan, Peter Brandt, Manasi Joshi, and Maheswaran Sathiamoorthy, *Google, Inc.*

Accelerating Large Scale Deep Learning Inference through DeepCPU at Microsoft 5
Minjia Zhang, Samyam Rajbandari, Wenhan Wang, Elton Zheng, Olatunji Ruwase, Jeff Rasley, Jason Li, Junhua Wang, and Yuxiong He, *Microsoft AI and Research*

MLOp Lifecycle Scheme for Vision-based Inspection Process in Manufacturing 9
Junsung Lim, Hoejoo Lee, Youngmin Won, and Hunje Yeon, *Samsung Research*

Shooting the moving target: machine learning in cybersecurity 13
Ankit Arun and Ignacio Arnaldo, *PatternEx*

Deep Learning Inference Service at Microsoft 15
Jonathan Soifer, Jason Li, Mingqin Li, Jeffrey Zhu, Yingnan Li, Yuxiong He, Elton Zheng, Adi Oltean, Maya Mosyak, Chris Barnes, Thomas Liu, and Junhua Wang, *Microsoft*

Handling Heterogeneity, Distribution, and Scale

Towards Taming the Resource and Data Heterogeneity in Federated Learning 19
Zheng Chai, *George Mason University;* Hannan Fayyaz, *York University;* Zeshan Fayyaz, *Ryerson University;* Ali Anwar, Yi Zhou, Nathalie Baracaldo, and Heiko Ludwig, *IBM Research–Almaden;* Yue Cheng, *George Mason University*

Measuring and Diagnosing Production ML

MPP: Model Performance Predictor 23
Sindhu Ghanta, Sriram Subramanian, Lior Khermosh, Harshil Shah, Yakov Goldberg, Swaminathan Sundararaman, Drew Roselli, and Nisha Talagala, *ParallelM*

Optimizing and Tuning

Low-latency Job Scheduling with Preemption for the Development of Deep Learning 27
Hidehito Yabuuchi, *The University of Tokyo;* Daisuke Taniwaki and Shingo Omura, *Preferred Networks, Inc.*

tensorflow-tracing: A Performance Tuning Framework for Production 31
Sayed Hadi Hashemi, *University of Illinois at Urbana-Champaign and National Center for Supercomputing Applications;* Paul Rausch; Benjamin Rabe, *University of Illinois at Urbana-Champaign and National Center for Supercomputing Applications;* Kuan-Yen Chou, *University of Illinois at Urbana-Champaign;* Simeng Liu, *University of Illinois at Urbana-Champaign and National Center for Supercomputing Applications;* Volodymyr Kindratenko, *National Center for Supercomputing Applications;* Roy H Campbell, *University of Illinois at Urbana-Champaign*

Disdat: Bundle Data Management for Machine Learning Pipelines 35
Ken Yocum, Sean Rowan, and Jonathan Lunt, *Intuit, Inc.;* Theodore M. Wong, *23andMe, Inc.*

TonY: An Orchestrator for Distributed Machine Learning Jobs 39
Anthony Hsu, Keqiu Hu, Jonathan Hung, Arun Suresh, and Zhe Zhang, *LinkedIn*

Transfer Learning for Performance Modeling of Deep Neural Network Systems 43
Md Shahriar Iqbal, *University of South Carolina;* Lars Kotthoff, *University of Wyoming;* Pooyan Jamshidi, *University of South Carolina*

Solutions and Platforms

KnowledgeNet: Disaggregated and Distributed Training and Serving of Deep Neural Networks 47
Saman Biookaghazadeh, Yitao Chen, Kaiqi Zhao, and Ming Zhao, *Arizona State University*

Continuous Training for Production ML in the TensorFlow Extended (TFX) Platform 51
Denis Baylor, Kevin Haas, Konstantinos Katsiapis, Sammy Leong, Rose Liu, Clemens Menwald, Hui Miao, Neoklis Polyzotis, Mitchell Trott, and Martin Zinkevich, *Google Research*

Katib: A Distributed General AutoML Platform on Kubernetes 55
Jinan Zhou, Andrey Velichkevich, Kirill Prosvirov, and Anubhav Garg, *Cisco Systems;* Yuji Oshima, *NTT Software Innovation Center;* Debo Dutta, *Cisco Systems*

Stratum: A Serverless Framework for the Lifecycle Management of Machine Learning-based Data Analytics Tasks 59
Anirban Bhattacharjee, Yogesh Barve, Shweta Khare, Shunxing Bao, and Aniruddha Gokhale, *Vanderbilt University;* Thomas Damiano, *Lockheed Martin Advanced Technology Labs*