MLModelScope: Evaluate and Profile ML Models at Scale and Across Stack

Cheng Li*, Abdul Dakkak*, Jinjun Xiong†, Wen-Mei Hwu*
{cli99, dakkak, w-hwu}@illinois.edu, jinjun@us.ibm.com
*University of Illinois Urbana-Champaign, †IBM Research Yorktown

Motivation
- The current landscape of ML is rife with diverse models, HW/SW stacks, and evaluation methodologies
- ML model performance is impacted by the interplay between frameworks, system libraries, compilers, and hardware platforms
- Current evaluation is laborious and error-prone and there is a lack of tools that make it fair and simple to compare different ML innovations
- Enables understanding ML model performance at each level of the HW/SW stack

Modular Design
- The system design is flexible enough to be extended to accommodate generic ML pipelines
- Makes it easy and simple to compare different ML innovations

Evaluation at Scale
- Key distributed design components:
  - Common prediction interface which works for any framework and model predictors
  - Profilers and Tracers for across stack profiling

Profiling Across Stack
- To introspect model performance across the HW/SW stack, currently researchers have to switch between tools and manually stitch the outputs (might not be possible)
- A scalable across-stack profiling scheme that correlates and aggregates profiles from different profiling providers into a single timeline
- Automatic model performance analysis, characterization, and reporting pipeline
- While we currently focus on ML model performance on GPUs, the across-profiling design is general and extensible

Case Study: MLPerf Inference ResNet50 v1.5
Characterization of MLPerf_ResNet50_v1.5 in NGC TensorFlow 19.06 on EC2 P3 (V100 GPU)

MLModelScope: Evaluate and Profile ML Models at Scale and Across Stack

An open-source, extendable and customizable framework to evaluate and profile ML models at scale and across stack
- Command line, API or web interface
- End-to-end profiling at different abstraction levels
- Built-in support for Caffe, Caffe2, CNTK, MXNet, PyTorch, TensorFlow, and TensorFlow RT
- Runs on X86, PPC, ARM using CPU, GPU, and FPGA
- An online portal of continuously updated evaluation and profiling results

Resources
- Documentation at docs.mlmodelscope.org
- Learn more about the center’s work at C3SR.com