MLModelScope: Evaluate and Measure ML Models

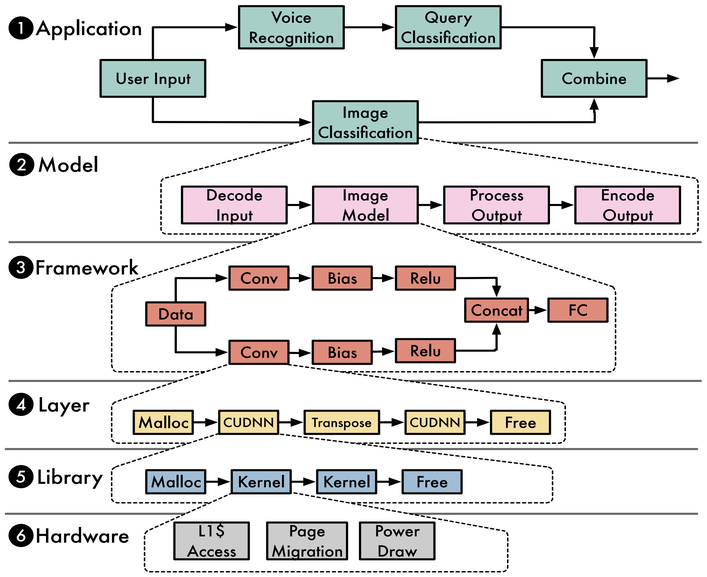
Abdul Dakkak*, Cheng Li*, Jinjun Xiong†, Wen-Mei Hwu*
{dakkak,cli99,w-hwu}@illinois.edu, jinjun@us.ibm.com
*University of Illinois Urbana-Champaign, †IBM Research Yorktown

C3SR center for cognitive computing systems research

IBM | ILLINOIS

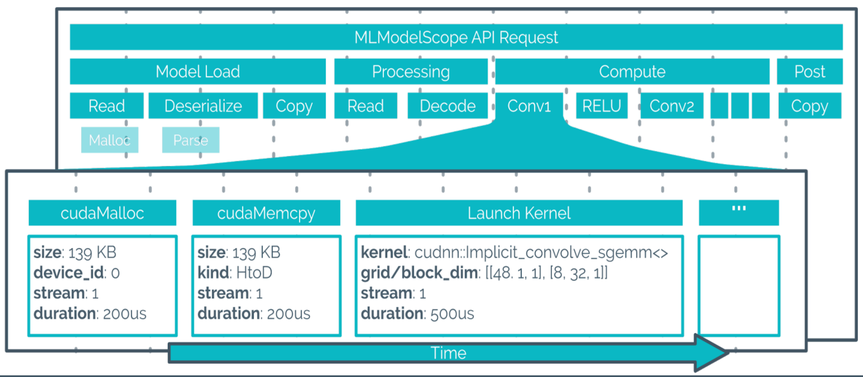
Motivation

- The current landscape of ML is rife with incompatible frameworks, models, evaluation methodologies, and system stacks
- AI applications are complicated where pipelines leverage models, frameworks, libraries, and HW
- Evaluating and comparing the benefits of proposed AI innovations is both arduous and error prone



Currently, the community lacks of a **standard tool** that:

- enables understanding of the proposed models and systems at each level of SW/HW stacks
- makes it simple to reproduce, evaluate, debug, compare, and manage reported results
- allows users to measure proposed models and systems within one's own AI workflow



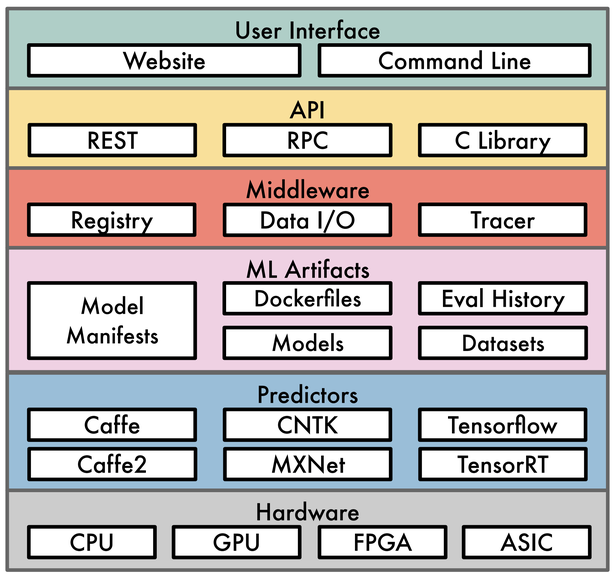
MLModelScope

An open source, batteries-included and customizable platform to aid users in model **evaluation** and **introspection** across datasets, frameworks, and systems.

- Application developers can discover and experiment with models that are applicable to their problems
- Data scientists can design and optimize models with deployment and hardware in mind
- System designers can profile model execution at all abstraction levels to tailor SW/HW stacks for the latest ML workloads

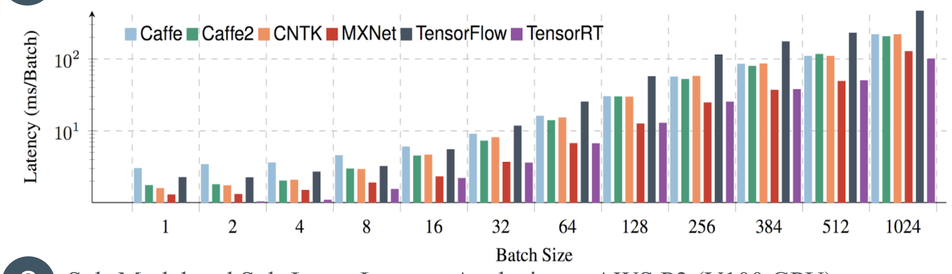
Current Capabilities

- An online public portal of continuously updated assets, evaluation results, and access to hardware resources
- Built-in support for evaluating and profiling models in TensorRT, TensorFlow, PyTorch, MXNet, Caffe, Caffe2, and CNTK
- Works on X86, PPC, ARM using CPU, GPU, and FPGA
- Can be used as an application with a web, command line, or API interface or can be compiled into a standalone library
- End-to-end profiling at different abstraction levels --- web API calls, model layer, GPU kernel and CUDA execution profile, hardware performance counter
- Top1/Top5 accuracy, model divergence analysis, static analysis such as theoretical flops calculation and memory requirement
- Evaluation and profiling report generation

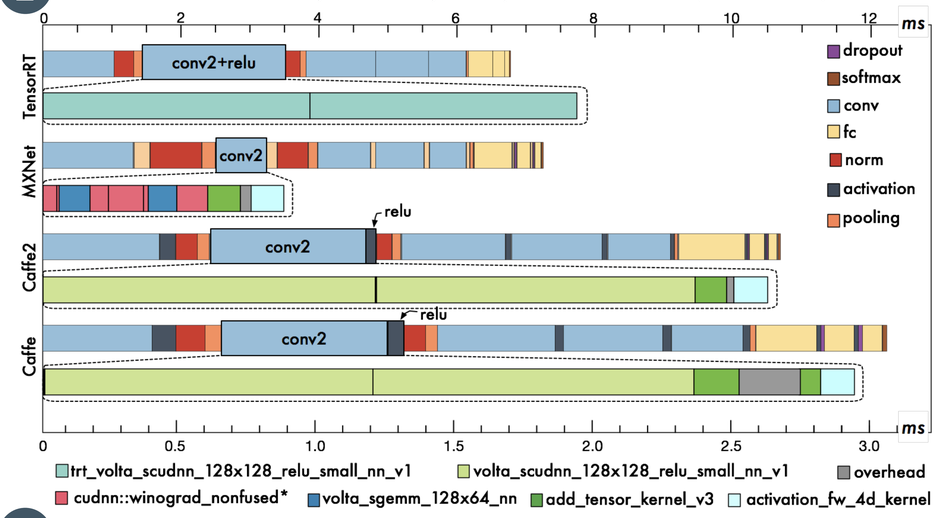


Case Study: In-depth Analysis of AlexNet

1 Model Latency across Frameworks on AWS P3 (V100 GPU)



2 Sub-Model and Sub-Layer Latency Analysis on AWS P3 (V100 GPU)



3 Model Latency across Systems

Future Work

- Adding more models, frameworks, data sets, and hardware
- Advising application developers on the best model given their dataset, target accuracy, $ budget, and target latency
- Profiling and analyzing performance on future and/or hypothetical hardware
- Providing similar levels of support for model training
- Support for hosting competitions, scoreboard and publication reproducibility

MLModelScope.org