Faster scalable ML model deployment using ONNX and open source tools

Faith Xu
Senior Program Manager, Machine Learning Platform
Microsoft
San Francisco, USA
faxu@microsoft.com

2020 IEEE Infrastructure Conference | 978-1-7281-7728-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IEEECONF47748.2020.9377615

***Abstract*—As ML developments shift from research to real world, we encounter many deployment challenges. Teams may be experimenting with various training frameworks, with deployments targeting multiple platforms and hardware. While training using one framework with one hardware target can easily be managed, it becomes challenging with a matrix of multiple frameworks and deployment targets. This fragmented ecosystem introduces deployment complexities and oftentimes custom code is needed to maximize performance for each scenario, which is time-consuming to maintain when models are updated.**

To streamline this, the interoperable ONNX model format and ONNX Runtime inference engine can be utilized to deploy models performantly across a variety of hardware. Models trained from PyTorch, Tensorflow, scikit-learn, CoreML, and more can all be converted to the common ONNX format, and the model can then be inferenced using the cross-platform performance-focused ONNX Runtime inference engine, which supports various hardware options for acceleration across CPU and GPUs.

ONNX Runtime is already used in key Microsoft services, on average realizing 2x performance improvements. In this session, we share an overview of ONNX Runtime, success stories and usage examples from high volume product groups at Microsoft, and demonstrate ways to integrate this into your AI workflows for immediate impact.

Keywords*—*infrastructure, machine learning, ONNX