What's new in ONNX RUNTIME

v1.10 (December 2021)

v1.11 (March 2022)

v1.12 (July 2022)

Ryan Hill | Software Engineer, Microsoft
ryanhill@microsoft.com

ONNX Runtime overview

Performant runtime engine for ONNX models
- Inference acceleration
- Training acceleration

Cross platform, cross architecture
- Windows/Linux/Mac (X86, X64, ARM)
- Mobile (iOS/Android)
- Web

Cross language
- Python
- C/C++
- C#
- Java
- JavaScript
- Objective-C

ONNX Runtime production adoption

Within Microsoft:

- 160+ models, 2.5x performance improvement
- 1 trillion+ daily inferences

Windows · Bing · Office 365 · Visual Studio Code · ML.NET · Power BI

Azure Cognitive Services · Microsoft Advertising · Microsoft Teams · skype · PowerApps

Azure Synapse / SQL Server · Azure ML · Azure Media Services · Azure Kinect DK · Azure Stack Edge · Surface

Community:

Adobe · ANT GROUP · ATLAS EXPERIMENT · GhostWriter · Hugging Face · hypefactors

intel · navitaire an AMADEUS company · NVIDIA · ORACLE · PEAKSPEED

PTW THE DOSIMETRY COMPANY · Rockchip 瑞芯微电子 · samtec · sas THE POWER TO KNOW. · TOPAZ LABS · UNREAL ENGINE

USDA · vespa · WRITER · XILINX

New Features

Use ops as a math library [1.12]

- No more copy & paste of internal code
- `OrtApi::CreateOp`
- `OrtApi::InvokeOp`

Feed external initializers as byte arrays for model inferencing [1.12]

- Useful for large models
- Previously required file on disk, now can be done entirely in memory

Performance Improvements

Transpose optimizer [1.10]

- Push and cancel transpose ops
- Significantly improves performance for models requiring layout transformation

Small Size Optimizations [1.11]

- std::vector<> to TensorShape & InlinedVector
- Allocations dropped from 70 to 6
- One user scenario drops from 479µsec to 360µsec

Quantization

- CNN speedups (Times are in milliseconds)

Pixel 4	Big Core			Little Core		
Model	1.9	1.12	%	1.9	1.12	%
efficientnet	19.71	13.2	33%	119.8	63.2	47%
Mobilenet_edgetpu	20.2	17.7	12%	109.2	75.74	31%
Cartoongan	1383.46	940	32%	7078	5576.92	21%
Mobilenet_v2_128	2.27	1.99	12%	12.79	10.29	20%
Deeplabv3	265.5	234.9	12%	1761.88	869	51%

Execution Providers

CUDA

- Preview support for CUDA Graphs to remove CPU overhead associated with launching CUDA kernels sequentially [1.11]

TensorRT

- 8.2.3 support [1.11]

OpenVINO

- 2022.1.0 support [1.11]

DirectML

- Updated additional operator coverage for newer opset
- support full precision uint64/int64 for 48 operators [1.10]

SNPE (Qualcomm)

- Enables accelerated inference execution on Snapdragon, Adreno, and Hexagon products.

ORT Mobile

NHWC/NCHW conversion at runtime

- Num Samples, Height, Width, Channels
- Kernels that prefer one layout "just work"
- Conversions can be optimized away in many cases

C# cross platform support for Android and iOS

- Xamarin
- .NET 6/MAUI [COMING SOON]

Android and iOS packages in **full** ORT builds

- Use an ONNX model with all ops/types that ORT supports
- Trade off of larger binary size

ORT Web

New WebAssembly core [1.10]

- Previously was a separate JavaScript project (onnxjs)
- Now build option to create ORT WebAssembly static library [1.11]
- Faster, uses less memory, and smaller footprint

Initial XNNPACK support [1.12]

OpenGL support (performance-contingent) [1.12]

ONNX Runtime Extensions

What is it?

- A library of custom ops for common pre/post processing
- Useful for NLP, vision, text domains
- Converter tool

Where is it?

- https://github.com/microsoft/onnxruntime-extensions
- Included with ORT through static library build option

What's next?

We will continue to

- Keep updating with ONNX
- Keep working with our partners on Execution Providers
- Keep listening to the community

Check us out on the web:

- https://onnxruntime.ai
- https://www.youtube.com/onnxruntime
- https://twitter.com/onnxruntime
- https://www.linkedin.com/company/onnxruntime