ONNX | Meetup June 24 2022

ONNX Steering Committee

- Prasanth Pulavarthi (MSFT)
- Alexander Eichenberger (IBM)
- Mayank Kaushik (NVIDIA)
- Rajeev Nalawadi (Intel)
- Andreas Fehlner (TRUMPF Laser GmbH)

ONNX

Welcome!

Open Neural Network Exchange

The open standard for machine learning interoperability

GET STARTED

ONNX is an open format built to represent machine learning models. ONNX defines a common set of operators - the building blocks of machine learning and deep learning models - and a common file format to enable AI developers to use models with a variety of frameworks, tools, runtimes, and compilers. LEARN MORE >

KEY BENEFITS

Interoperability

Develop in your preferred framework without worrying about downstream inferencing implications. ONNX enables you to use your preferred framework with your chosen inference engine.

SUPPORTED FRAMEWORKS >

Hardware Access

ONNX makes it easier to access hardware optimizations. Use ONNX-compatible runtimes and libraries designed to maximize performance across hardware.

SUPPORTED ACCELERATORS >

ONNX Community

ABBYY
aizon
Alibaba Group 阿里巴巴集团
AMD
arm
aws
Baidu 百度
BECKHOFF
BITMAIN
cadence
CEVA
Facebook Open Source
GRAPHCORE
habana
HAILO
Hewlett Packard Enterprise
HUAWEI
IBM
Idein Inc
intel
MathWorks
MAXAR
MEDIATEK
mi
Microsoft
NVIDIA
NXP
OctoML
ORACLE
OPEN AI LAB 开放智能
Preferred Networks
SIEMENS
SONY
Qualcomm
sas
商汤 sensetime
skymizer
SYNOPSYS
Tencent
unity
verizon media
vmware
WOLFRAM
Yandex
ZETANE

And more...

ONNX is a Community Project

Steering Committee
https://github.com/onnx/steering-committee

Prasanth Pulavarthi (Microsoft)
Alexander Eichenberger (IBM)
Mayank Kaushik (NVIDIA)
Rajeev Nalawadi (Intel)
Andreas Fehlner* (TRUMPF Laser)

Special Interest Groups (SIGs) and Working Groups
https://github.com/onnx/sigs

Architecture & Infra: Liqun Fu*, Ke Zhang

Operators: Michał Karzyński, Ganesan Ramalingam

Converters: Thiago Crepaldi*, Kevin Chen*

Model Zoo & Tutorials: Jacky Chen*

Pre-processing (WG): Joaquin Anton

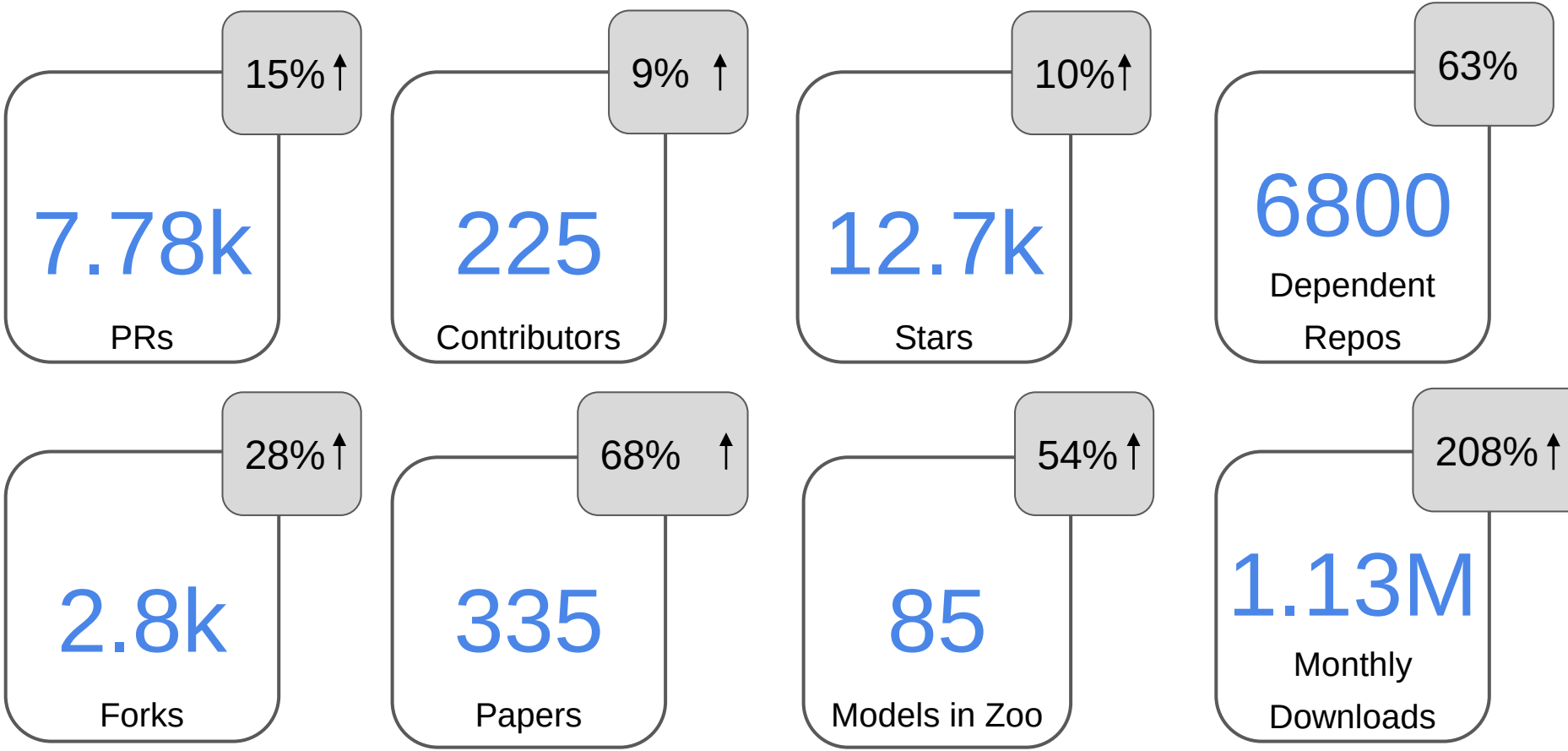
* = new to role since October 2021

ONNX

State of the Union

Tools and Companies that Support ONNX

Engagement & usage (from 10/20/21 to 06/13/2022)



ONNX

Release Update

ONNX 1.11 Released

Release v1.11.0 · onnx/onnx (github.com)

ONNX v1.11.0 comes with following updates:

- Opset 16 introduced with new and updated operators
- Added Model hub (to pull pre-trained models from zoo)
- Compose utilities to create combined model with preprocessing & inference
- Functionbuilder utility to help create function ops
- Bugfixes and infrastructure improvements
- Documentation updates

Visit the release page on GitHub for more details

Thank you everyone for your countless hours of work!

ONNX 1.12 Released

Release v1.12.0 · onnx/onnx (github.com)

ONNX v1.12 comes with following updates:

- Opset 17 introduced with new and updated operators
- Shape inference enhancements
- Bugfixes and infrastructure improvements
- Documentation updates
- Add Python 3.10 and drop Python 3.6 support
- Drop support for x86 (32-bit) Linux due to low usage

Visit the release page on GitHub for more details

Thank you everyone for your countless hours of work!

ONNX

ONNX Roadmap Process

ONNX roadmap discussions (Process)

12 Roadmap requests were selected for further progress & assigned SIG/WG's

ONNX

Roadmap
Requests
Current status

ONNX Roadmap Items (Status) - 1

Topics	Proposed SIG's	Current Status or Future positioning
New operators for data processing to cover ML pipeline – Nakaike (IBM)	Operators, Pre-processing	Identified from operators group about significant overlap with extensions already implemented. Work continues on processing dates (feature)
C API for C++ components of ONNX (to assist in wrapper for model checker functionality) – Pocock (Oracle)	Arch/Infra	C API to wrap protobuf which python, C#, Java can interact to emit ONNX models continues as long term goal. Currently C#, Java have their own construction and validation code
Better support for emitting ONNX models from other languages beyond Python – Pocock (Oracle)	Arch/Infra	"Same as above"
Add meta information in tensors – Croome (Greenwaves)	Arch/infra, Converters, Release	Identifying path to get more structured quantization information in onnx to match/exceed tflite's abilities.
E2E pipeline with ONNX operators (include Keras, TF, Scikit-learn/Spark pipeline preprocessing flows) using single graph – Sica (IBM)	Arch/Infra, Model Tutorial, Operators, Pre-processing	Identified as long term intercept, further refining the proposal
Converters improvement suggestions (tensorflow-onnx, Keras2Onnx) for better graph optimizations – Sica (IBM)	Converters, Operators	Higher functioning ops specifically targeting (LSTM/GRU in particular) have gotten support in tfonnx (converters). Efforts will continue as new opportunities rise

ONNX Roadmap Items (Status) - 2

Topics	Proposed SIGs	Current Status or Future Positioning
Address gaps with Opset conversions across broad set of models – Sabharwal (Intel)	Arch/Infra, Converters, Release	Past two ONNX releases have fixed subset of issues with Opset conversions/compatibility. Efforts to continue in future as newer Opsets get introduced
ONNX model zoo example for E2E distributed training scenario of large models – Esteves (Intel)	Model Tutorial	Identified as long term intercept
Define concept of federated learning for ONNX – Esteves (Intel)	Operators	Identified no new ONNX operators required, exploring further on solutions that are Framework/runtime agnostic
Improvements to shape inference implementation – McCarter (Lighmatter)	Arch/Infra	Submitter analyzing further on the Shape inference improvements to be targeted for future
Introduce ONNX model provenance & security to safeguard against manipulations – Karumanchi (Intel)	Arch/Infra, Model Tutorial, Pre-processing	Initial metadata fields defined to establish machine readable ONNX model provenance https://github.com/onnx/onnx/issues/3958
ONNX model zoo support for quantized and mixed precision models.– Karumanchi (Intel)	Model Tutorial, Operators	Related to ONNX model metadata field definition, will target couple of mixed precision models in zoo once Issue #3958 finalized

Call for Volunteers

- We need more maintainers across ONNX for managing PRs
- Architecture & Infra and Operators in particular seem to have a lot of pending PRs.
- Please volunteer!
- https://github.com/onnx/onnx/blob/main/community/sigs.md

Thank you ...

- Please stay engaged and continue your contributions to ONNX and its related projects.
- Remember to use the following ONNX resources:
  - Website: https://onnx.ai/
  - GitHub: https://github.com/onnx
  - Slack: (join https://slack.lfai.foundation - email, password, then find onnx-general)
  - Calendar: https://onnx.ai/calendar
  - Mailing List: https://lists.lfai.foundation/g/onnx-announce
  - Twitter: https://twitter.com/onnxai

Questions?