ONNX

(Open Neural Network Exchange)

Current status & beyond

Motivation: Lots of DL frameworks!

TensorFlow PyTorch mxnet GLUON

Keras Caffe2 CNTK Chainer

PaddlePaddle neon ML.NET And more!

- Create open standard format for deep learning models
- Provide deep learning model interoperability among frameworks
- Protobuff: github.com/onnx/onnx/blob/master/onnx/onnx.proto

Industry wide participation

At inception: aws, Facebook Open Source, Microsoft

Dec 2017

Industry partners: AMD, Hewlett Packard Enterprise, Tencent, intel AI, IBM, MEDIATEK, arm, HUAWEI, MathWorks, Qualcomm, NVIDIA, Oath:, unity, Baidu 百度, Alibaba Group 阿里巴巴集团, habana, Preferred Networks, BITMAIN, skymizer, mi, synopsys, NXP, CEVA, sas

Current Spec as of May 2019:

1.5

Current status: Framework & converter support

8 Frameworks

PyTorch, Caffe2, mxnet, ML.NET, PaddlePaddle, CNTK, MATLAB, Chainer

```
#Export Alexnet from PyTorch
import torch
torch.onnx.export(model, dummy_input, "alexnet.onnx",...)

#Import Alexnet to Caffe2
import onnx
import caffe2.python.onnx.backend as backend
model = onnx.load("alexnet.onnx")
rep = backend.prepare(model, device="CPU")
```

6 Converters

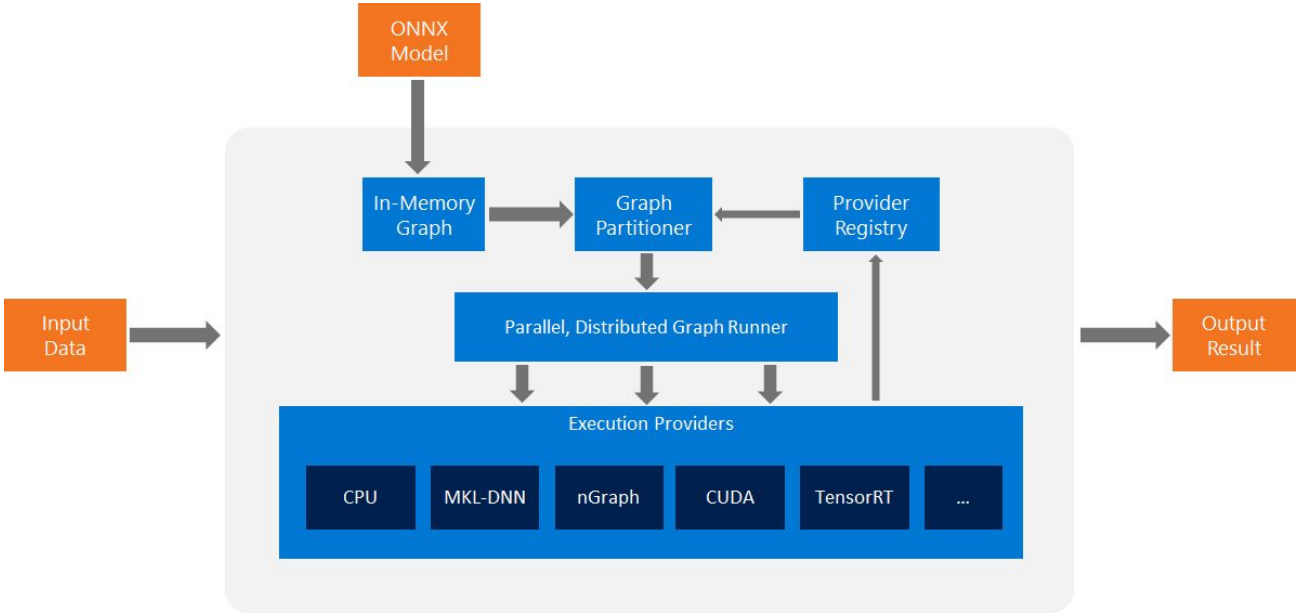
TensorFlow, Keras, LibSVM, dmlc XGBoost, scikit learn, ML

```
#To convert TF model to ONNX model
python -m tf2onnx.convert\
    --input tests/models/fc-layers/frozen.pb\
    --output tests/models/fc-layers/model.onnx\
```

Current status: Runtimes, Compilers, & Visualizers

11 Runtimes	ONNX RUNTIME, intel AI, Qualcomm, NVIDIA, SYNOPSYS, SOPHON, Tencent, Windows, vespa, habana, MACE Mobile AI Compute Engine
4 Compilers	CEVA, nGraph, skymizer, tvm
2 Visualizers	NETRON, Visual DL

Current status: ONNX Runtime



Will be open sourced at //BUILD conference (May 6)

On average 2X performance boost*

Supports quantization

More EPs in progress

https://azure.microsoft.com/en-us/blog/onnx-runtime-for-inferencing-machine-learning-models-now-in-preview/

Current status: ONNXIFI & ONNX.js

ONNXIFI

API to connect deep learning frameworks to backends. Implemented for Caffe2 (github.com/onnx/onnx/blob/master/docs/ONNXIFI.md)

ONNX.js

Javascript library for running ONNX models on browsers and on Node.js
Benchmark show 7 ~ 26X performance improvement over similar efforts (github.com/Microsoft/onnxjs)

```
<script src="https://cdn.jsdelivr.net/npm/onnxjs/dist/onnx.min.js"></script>
const myOnnxSession = new onnx.InferenceSession();
myOnnxSession.loadModel("./my-model.onnx")
```

Current status: Model zoo (github.com/onnx/models)

Image Classification

- MobileNet
- ResNet
- SqueezeNet
- VGG
- Bvlc_AlexNet
- Bvlc_GoogleNet
- Bvlc_reference_CaffeNet
- Bvlc_reference_RCNN_ILSVRC13
- DenseNet121
- Inception_v1
- Inception_v2
- ShuffleNet
- ZFNet512

Object Detection & Segmentation

- Tiny_YOLOv2
- SSD New!
- YOLO v3 New!
- DUC

Body, Face & Gesture Analysis

- ArcFace
- Emotion FerPlus

CNN models

Building blocks for RNN available

Current status: ONNX 1.5 spec

- Opset 10 adds operators to support object detection models such as Yolo v3, Faster RCNN, and SSD. Models will be added to the ONNX Model Zoo
- Quantization support (with first set of operators)
- Promote ONNX Function to support composing operators (support of more operators from other frameworks while limiting new operators)
- All experimental ops are removed and deprecated

Areas of development (gitter.im/onnx/Lobby)

- **Quantization:** Few more quantized ops to be added in the future (Intel)
- **Training:** Use cases defined. Technical discussion on enabling training IR, loss function & optimizer. Enabling ONNX Runtime for training (IBM)
- **Model zoo:** Implementing CI to automate compliance check (Huawei)
- **Mobile & Edge:** defining profiles for edge and associated use cases (Qualcomm)
- **Dataflow:** Define preprocessing op definitions. Reference implementation to be provided (NVidia)

New governance structure for wider participation

Transparent decision process & better technical decisions
github.com/onnx/onnx/tree/master/community

Community

- Members
- Contributors
- Approvers
- Member companies

Structure

- Steering Committee
- Special Interest Groups — Persistent
- Working groups — Temporary

Engagement & usage (Dec 2018 => Mar 2019)

10%↑

1169

Pull requests

15%↑

111

Contributors

16%↑

5658

Github stars

36%↑

241

Github repos

73%↑

40

Published Papers

Thank you!

onnx.ai