ONNX
(Open Neural Network Exchange)
Current status & beyond
Motivation: Lots of DL frameworks!

- 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
- Alibaba Group
- Habana
- Preferred Networks
- BITMAIN
- Skymizer
- Synopsys
- NXP
- CEVA
- SAS

Current Spec as of May 2019:
1.5
Current status: Framework & converter support

<table>
<thead>
<tr>
<th>8 Frameworks</th>
<th>6 Converters</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyTorch</td>
<td>TensorFlow</td>
</tr>
<tr>
<td>Caffe2</td>
<td>Keras</td>
</tr>
<tr>
<td>mxnet</td>
<td>LibSVM</td>
</tr>
<tr>
<td>PaddlePaddle</td>
<td>dmlc</td>
</tr>
<tr>
<td>CNTK</td>
<td></td>
</tr>
<tr>
<td>MATLAB</td>
<td></td>
</tr>
<tr>
<td>Chainer</td>
<td></td>
</tr>
</tbody>
</table>

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

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

#To convert TF model to ONNX model
```bash
python -m tf2onnx.convert\
  --input tests/models/fc-layers/frozen.pb\
  --output tests/models/fc-layers/model.onnx\n```
Current status: Runtimes, Compilers, & Visualizers

11 Runtimes
- ONNX Runtime
- Intel AI
- Qualcomm
- NVIDIA
- Synopsys
- Sophon
- Tencent
- Windows
- vespa
- Habana
- MACE

4 Compilers
- CEVA
- nGraph

2 Visualizers
- NETRON
- Visual DL
Current status: ONNX Runtime

- ONNX Model
- In-Memory Graph
- Graph partitioner
- Provider Registry
- Parallel, Distributed Graph Runner
- Execution Providers (CPU, MKL-DNN, nGraph, CUDA, TensorRT, ...)
- Output Result

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

- On average 2X performance boost*
- Supports quantization
- More EPs in progress

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](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](github.com/Microsoft/onnxjs))

```javascript
<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
- YOLO v3
- DUC
- New!

Body, Face & Gesture Analysis
- ArcFace
- Emotion FerPlus

Building blocks for RNN available

CNN models
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
- Working groups

- Persistent
- Temporary
Engagement & usage (Dec 2018 => Mar 2019)

- Pull requests: 1169 (10%↑)
- Contributors: 111 (15%↑)
- Github stars: 5658 (16%↑)
- Github repos: 241 (36%↑)
- Published Papers: 40 (73%↑)
Thank you!

onnx.ai