Microsoft

ONNX & ONNX Runtime

Weixing Zhang

Microsoft AI Frameworks

Agenda

- ❑ What is ONNX
- ❑ ONNX @ Microsoft
- ❑ What is ONNX Runtime
- ❑ How to create ONNX models

Open and Interoperable AI

ONNX

Open Neural Network Exchange

Open format for ML models

github.com/onnx

ONNX Partners

Alibaba Group 阿里巴巴集团 | AMD | arm | aws | Baidu 百度 | BITMAIN

CEVA | Facebook Open Source | GRAPHCORE | habana | Hewlett Packard Enterprise | HUAWEI

IBM | Idein Inc. | intel AI | MathWorks | MEDIATEK | Microsoft

Neural Network Libraries | mi | NVIDIA | NXP | Oath: A Verizon company | Preferred Networks

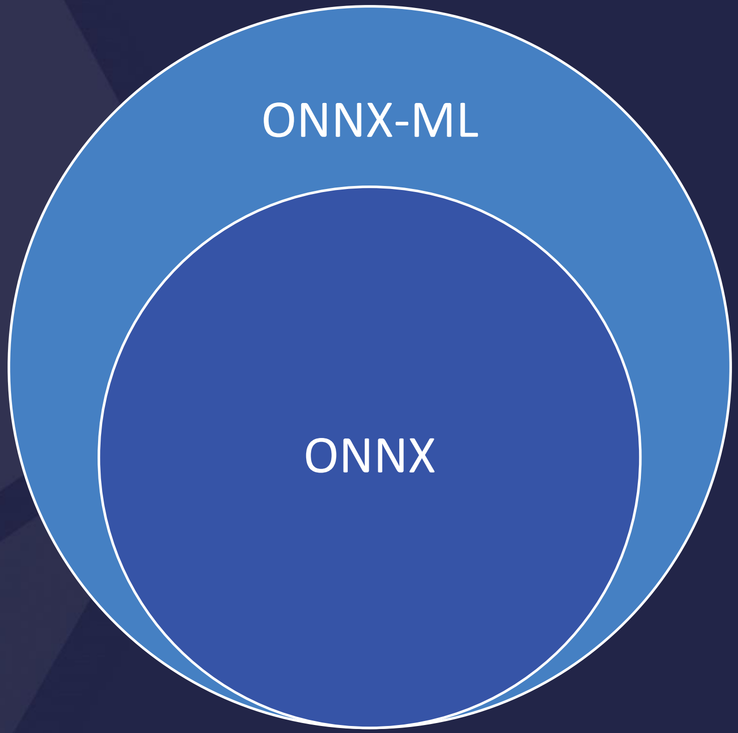
Qualcomm | sas | skymizer | SYNOPSYS | Tencent | unity

Key Design Principles

- Support DNN but also allow for traditional ML
- Flexible enough to keep up with rapid advances
- Compact and cross-platform representation for serialization
- Standardized list of well defined operators informed by real world usage

ONNX Spec

- File format
- Operators



File format

Model

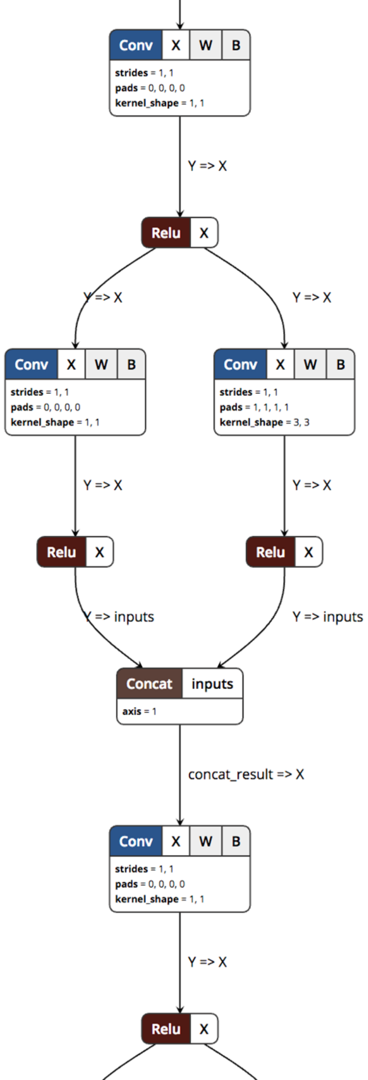
- Version info
- Metadata
- Acyclic computation dataflow graph

Graph

- Inputs and outputs
- List of computation nodes
- Graph name

Computation Node

- Zero or more inputs of defined types
- One or more outputs of defined types
- Operator
- Operator parameters



Data types

- Tensor type
  - Element types supported:
    - int8, int16, int32, int64
    - uint8, uint16, uint32, uint64
    - float16, float, double
    - bool
    - string
    - complex64, complex128
- Non-tensor types:
  - Sequence
  - Map

```
message TypeProto {
  message Tensor {
    optional TensorProto.DataType elem_type = 1;
    optional TensorShapeProto shape = 2;
  }
  // repeated T
  message Sequence {
    optional TypeProto elem_type = 1;
  };
  // map<K,V>
  message Map {
    optional TensorProto.DataType key_type = 1;
    optional TypeProto value_type = 2;
  };

  oneof value {
    Tensor tensor_type = 1;
    Sequence sequence_type = 4;
    Map map_type = 5;
  }
}
```

Operators

An operator is identified by <name, domain, version>

Core ops (ONNX and ONNX-ML)

- Should be supported by ONNX-compatible products
- Generally cannot be meaningfully further decomposed
- Currently 124 ops in ai.onnx domain and 18 in ai.onnx.ml
- Supports many scenarios/problem areas including image classification, recommendation, natural language processing, etc.

Custom ops

- Ops specific to framework or runtime
- Indicated by a custom domain name
- Primarily meant to be a safety-valve

Relu

Relu takes one input data (Tensor) and produces one output data (Tensor) where the rectified linear function, y = max(0, x), is applied to the tensor elementwise.

Version

This version of the operator has been available since version 6 of the default ONNX operator set. Other versions of this operator: Relu-1

Inputs

`X` : *T*
  Input tensor

Outputs

`Y` : *T*
  Output tensor

Type Constraints

`T` : *tensor(float16), tensor(float), tensor(double)*
  Constrain input and output types to float tensors.

Examples

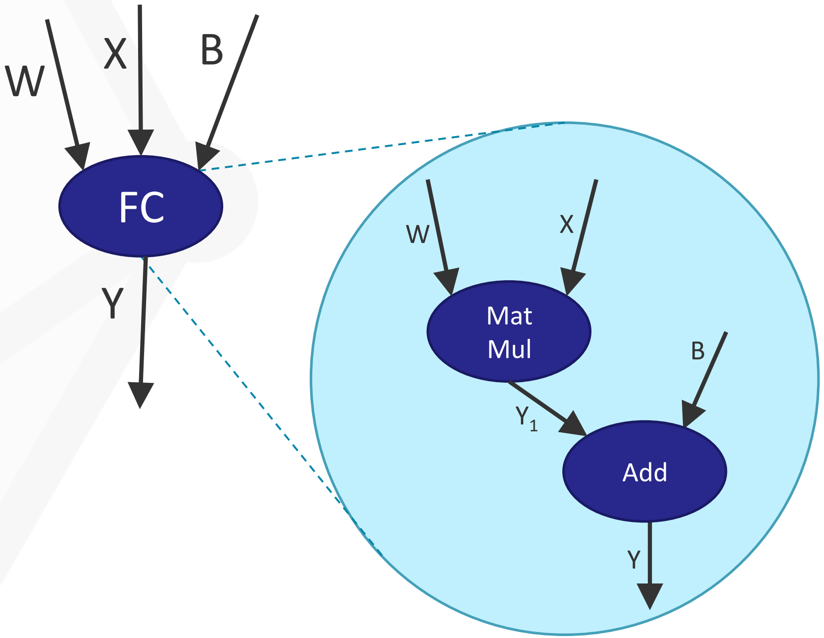
▼ relu

```
node = onnx.helper.make_node(
    'Relu',
    inputs=['x'],
    outputs=['y'],
)
x = np.random.randn(3, 4, 5).astype(np.float32)
y = np.clip(x, 0, np.inf)

expect(node, inputs=[x], outputs=[y],
       name='test_relu')
```

Functions

- Compound ops built with existing primitive ops

- Runtimes/frameworks/tools can either have an optimized implementation or fallback to using the primitive ops



Agenda

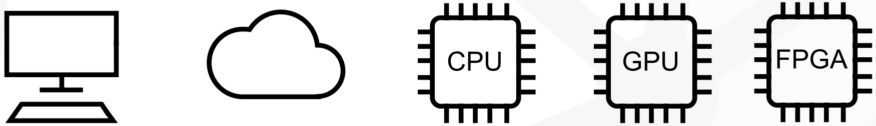
- ✓ What is ONNX
- ❑ ONNX @ Microsoft
- ❑ What is ONNX Runtime
- ❑ How to create ONNX models

ML @ Microsoft

- LOTS of internal teams and external customers
- LOTS of models from LOTS of different frameworks

Cognitive Toolkit · K · scikit learn · PyTorch · Caffe

- Different teams/customers deploy to different targets



ONNX @ Microsoft

PLATFORMS

AzureML WinML ML.Net

PRODUCTS

Bing
Microsoft Cognitive Services
Office 365
Power BI

Up to 14.6x

Performance gains seen by Microsoft services

100s of Millions

# of devices where ONNX Runtime is running

Billions

# of requests handled by ONNX Runtime across Microsoft services

ONNX @ Microsoft

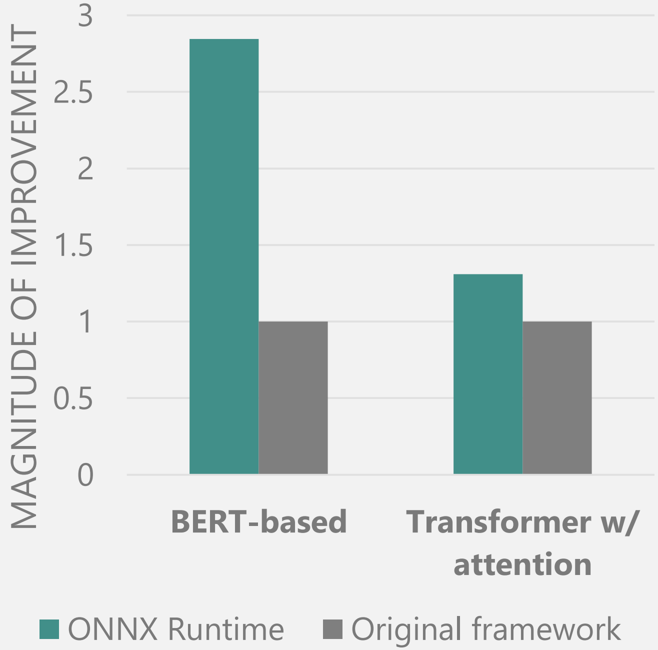
Bing QnA - List QnA and Segment QnA

- Two models used for generating answers
- Up to 2.8x perf improvement with ONNX Runtime

Query: empire earth similar games

PERFORMANCE

Up to **2.8x** perf improvement with ONNX Runtime

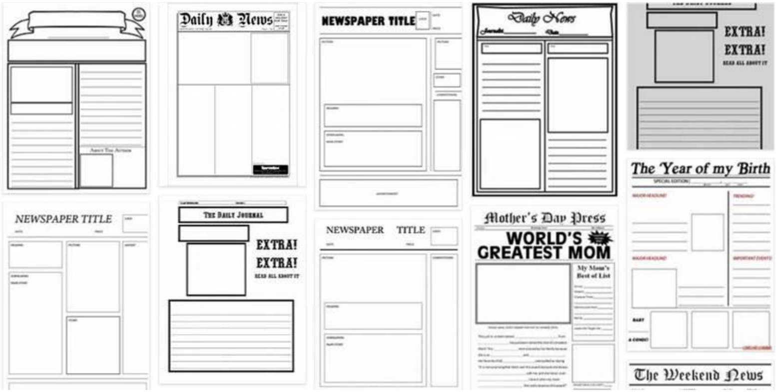


ONNX @ Microsoft

Bing Multimedia - Semantic Precise Image Search

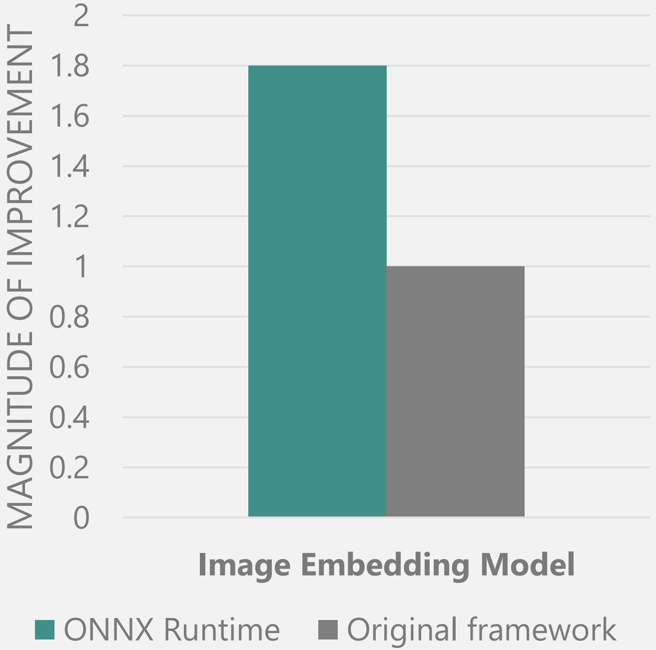
- Image Embedding Model - Project image contents into feature vectors for image semantic understanding
- 1.8x perf gain by using ONNX and ONNX Runtime

Query: newspaper printouts to fill in for kids



PERFORMANCE

<u>1.8x</u> perf improvement with ONNX Runtime



Agenda

- ✓ What is ONNX
- ✓ ONNX @ Microsoft
- ❑ What is ONNX Runtime
- ❑ How to create ONNX models

ONNX RUNTIME

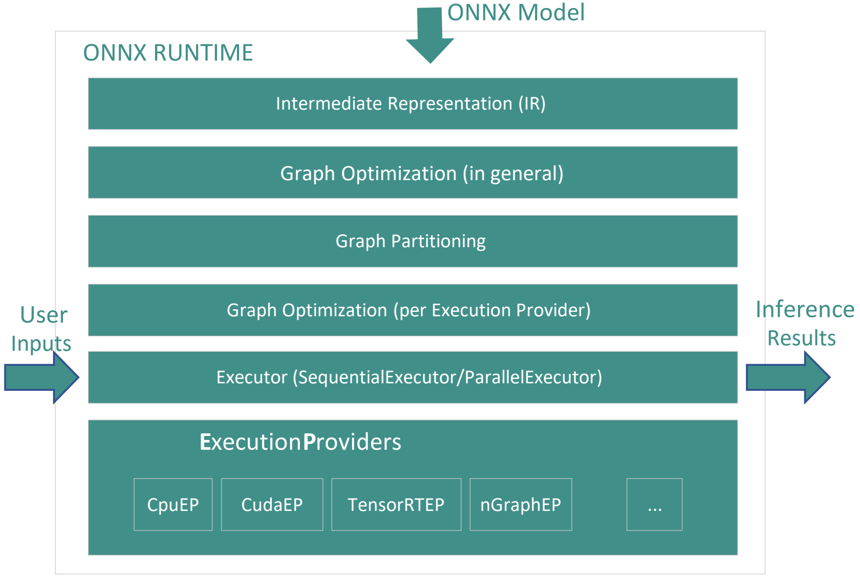
github.com/microsoft/onnxruntime

- High performance
- Cross platform
- Lightweight & modular
- Extensible

ONNX Runtime

- High performance runtime for ONNX models
- Extensible architecture to plug-in optimizers and hardware accelerators
- Supports full ONNX-ML spec (v1.2 and higher, currently up to 1.5)
- Works on Mac, Windows, Linux (ARM too)
- CPU, GPU, Intel edge devices, Nvidia Jeston Nano, …
- Python, C#, and C APIs
- Code Generation
- Training

ONNX Runtime – Architecture



Graph Optimization

- Node elimination (dropout, identity, etc.)
- Node fusion, constant folding, etc.

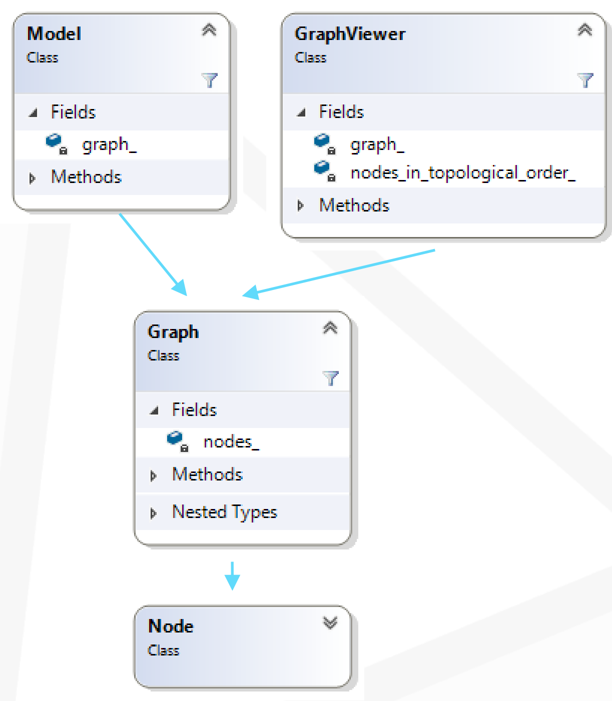
Graph Partitioning

- Graph partitioning based on execution providers' capability
- Greedy algo based on user preferences

Execution Provider

- Plug-in hardware accelerator
- Key APIs
  - GetCapability – given a graph, return a collection of sub-graphs it can run
  - Compile – given a sub-graph (node), return function pointers to run the sub-graph

ONNX Runtime – IR



Model/Graph/Node

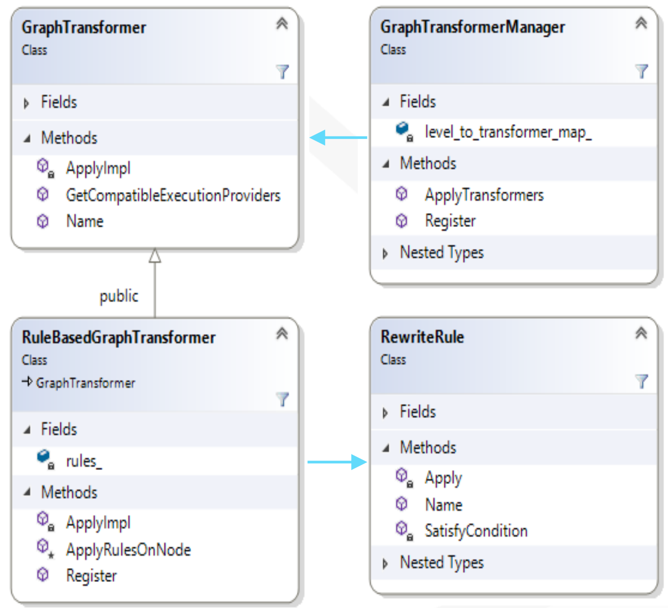
- In-memory object mapping to ONNX model file format design.
- Offering APIs to read/write a computational graph.

GraphViewer

Read-only view of a computational graph. Used in:

- IExecutionProvider (API between Runtime and hardware accelerator)
- Model evaluation (after model optimization and partitioning)

ONNX Runtime – Graph Optimization



RewriteRule

- An interface created for finding patterns (with specific nodes) and applying rewriting rules against a sub-graph.

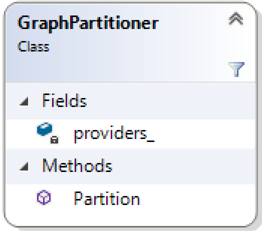
GraphTransformer

- An interface created for applying graph transformation with full graph editing capability.

TransformerLevel

- Level 0: Transformers anyway will be applied after graph partitioning (e.g. cast insertion, mem copy insertion)
- Level 1: General transformers not specific to any specific execution provider (e.g. drop out elimination)
- Level 2: Execution provider specific transformers

ONNX Runtime – Graph Partitioning



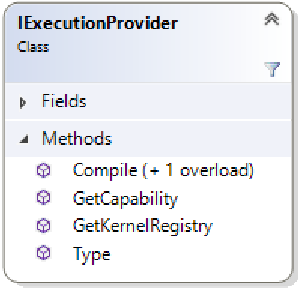
GraphPartitioner

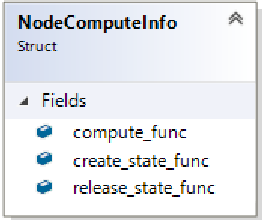
- Given a mutable graph, graph partitioner assigns graph nodes to each execution provider per their capability and idea goal is to reach best performance in a heterogeneous environment.
- ONNX RUNTIME uses a “greedy” node assignment mechanism
- Users specify a preferred execution provider list in order
- ONNX RUNTIME will go thru the list in order to check each provider’s capability and assign nodes to it if it can run the nodes.

FUTURE:

- Profiling based partitioning
- ML based partitioning

ONNX Runtime – Execution Provider





IExecutionProvider

A hardware accelerator interface to query its capability and get corresponding executables.

1) Kernel based execution providers
   These execution providers provides implementations of operators defined in ONNX (e.g. CPUExecutionProvider, CudaExecutionProvider, MKLDNNExecutionProvider, etc.)
2) Runtime based execution providers
   These execution providers may not have implementations with the granularity of ONNX ops, but it can run whole or partial ONNX graph. Say, it can run several ONNX ops (a sub-graph) together with one function it has (e.g. TensorRTExecutionProvider, nGraphExecutionProvider, etc.)

NodeComputeInfo

A data structure carries executables returned by runtime-based execution providers.

ONNX Runtime – API Example

```python
import onnxruntime

session = onnxruntime.InferenceSession("mymodel.onnx")

results = session.run([], {"input": input_data})
```

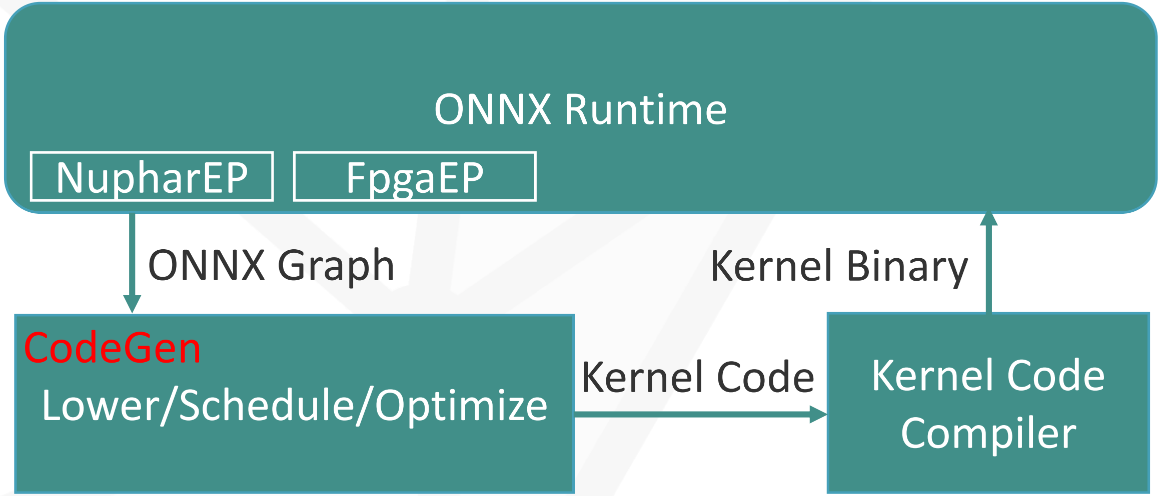
```csharp
using Microsoft.ML.OnnxRuntime;

var session = new InferenceSession("model.onnx");

var results = session.Run(input);
```

C#

…… also available for C

ONNX Runtime – CodeGen

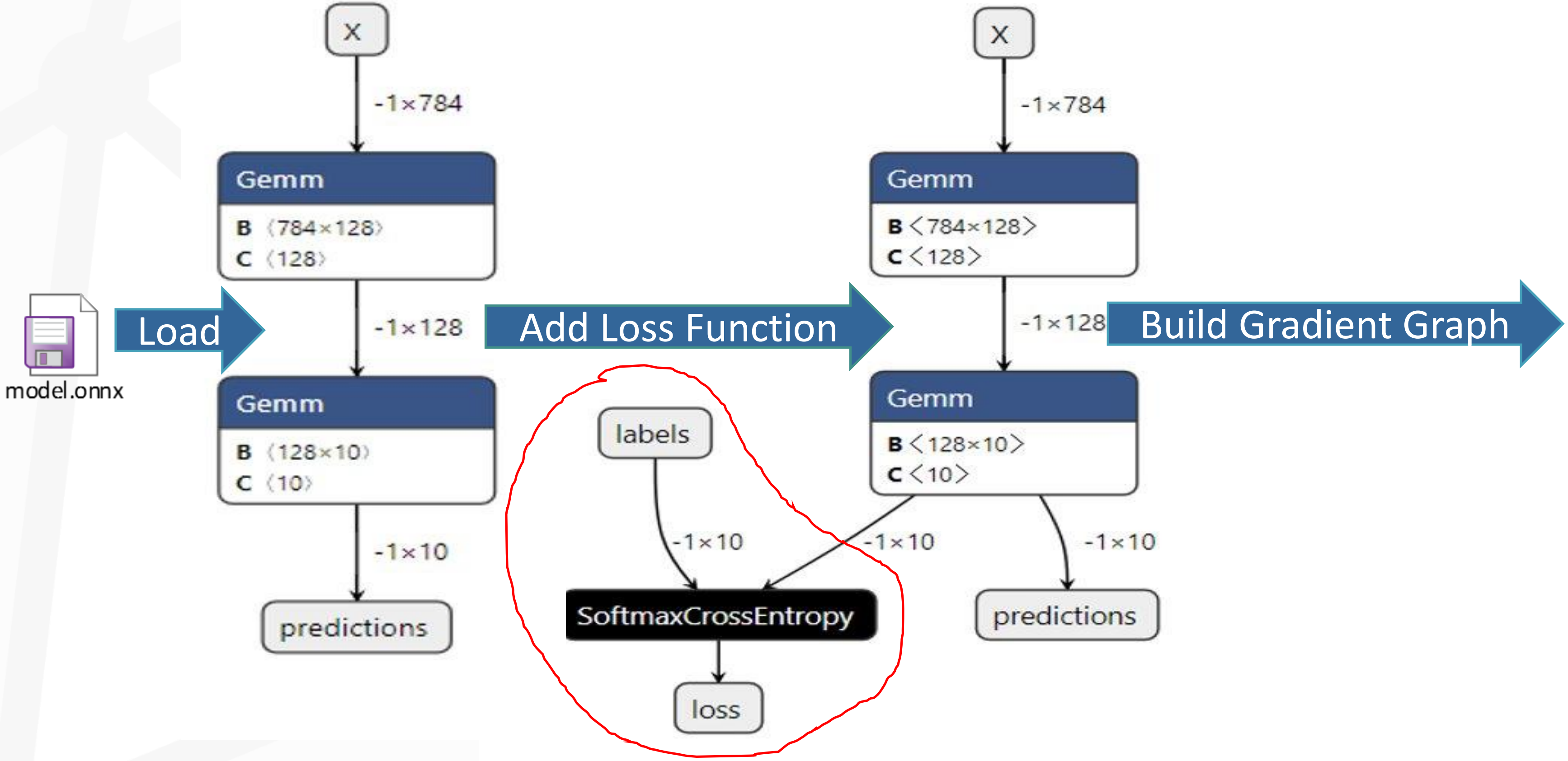
- TVM/Halide Based
- Used by NupharEP and FpgaEP
- NupharEP is open source

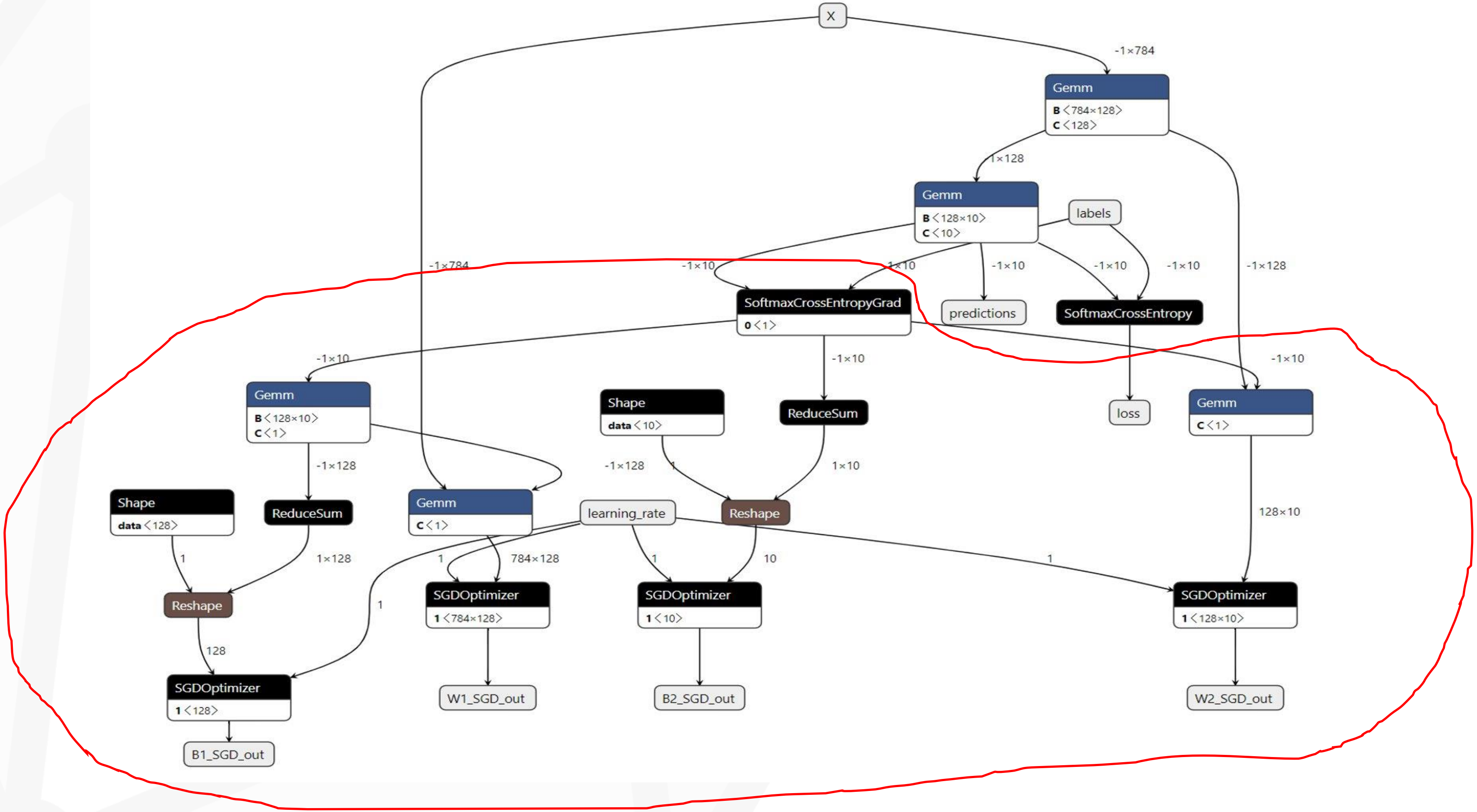


ONNX Runtime – Training

- Enables training on Device
  - Helps reinforcement learning, quantized retraining, …
  - A fast, light-weight runtime with both training and inferencing capability
    - ONNX RT is 3MB binary size, ONNX + Training about 5MB
- Enables large-scale training for multiple frontends and backends
  - A single, unified software stack that
    - Supports multiple training framework frontends (TensorFlow, PyTorch,…)
    - Supports multiple accelerator backends (GPU, …)
  - A combined SW and HW stack

ONNX Runtime – Training





X
-1×784
Gemm
B〈784×128〉
C〈128〉
1×128
Gemm
B〈128×10〉
C〈10〉
labels
-1×784
-1×10
1×10
-1×10
-1×10
-1×10
-1×128
SoftmaxCrossEntropyGrad
0〈1〉
predictions
SoftmaxCrossEntropy
-1×10
-1×10
-1×10
Gemm
B〈128×10〉
C〈1〉
Shape
data〈10〉
ReduceSum
loss
Gemm
C〈1〉
-1×128
-1×128
1×10
Shape
data〈128〉
ReduceSum
Gemm
C〈1〉
learning_rate
Reshape
128×10
1
1×128
1
784×128
1
10
1
Reshape
1
SGDOptimizer
1〈784×128〉
SGDOptimizer
1〈10〉
SGDOptimizer
1〈128×10〉
128
SGDOptimizer
1〈128〉
W1_SGD_out
B2_SGD_out
W2_SGD_out
B1_SGD_out

Agenda

- ✓ What is ONNX
- ✓ ONNX @ Microsoft
- ✓ What is ONNX Runtime
- ❑ How to create ONNX models

4 ways to get an ONNX model

- ONNX Model Zoo
- Services like Azure Custom Vision
- Convert existing models
- Train models in systems like Azure Machine Learning service

ONNX Model Zoo: github.com/onnx/models

Image Classification

This collection of models take images as input, then classifies the major objects in the images into a set of predefined classes.

Model Class	Reference	Description
MobileNet	Sandler et al.	Efficient CNN model for mobile and embedded vision applications. Top-5 error from paper - ~10%
ResNet	He et al., He et al.	Very deep CNN model (up to 152 layers), won the ImageNet Challe… Top-5…
SqueezeNet	Iandola et al.	A ligh… fewer… Top-5…
VGG	Simonyan et al.	Deep… Challe… Top-5…

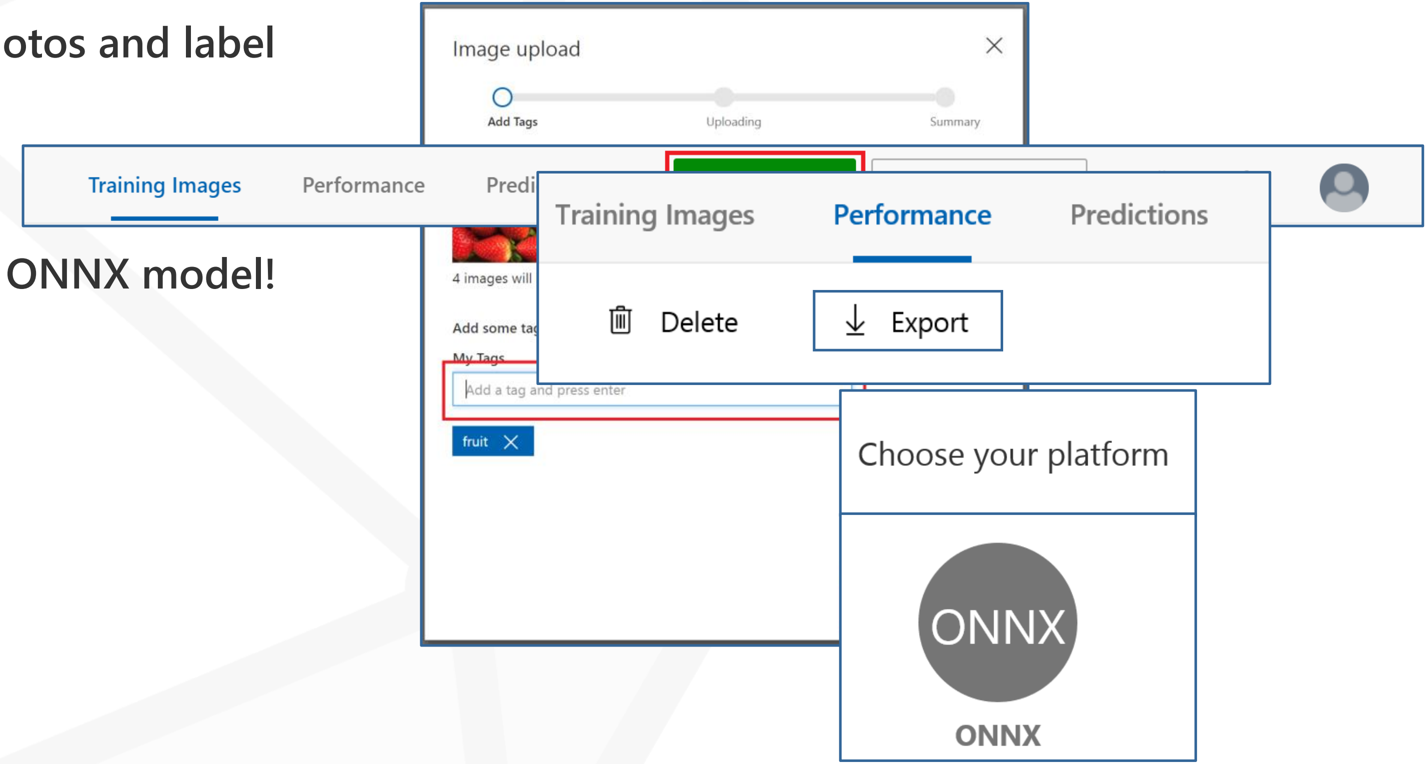
Model	Download	Checksum	Download (with sample test data)	ONNX version	Opset version	Top-1 accuracy (%)	Top-5 accuracy (%)
ResNet-18	44.6 MB	MD5	42.9 MB	1.2.1	7	69.70	89.49
ResNet-34	83.2 MB	MD5	78.6 MB	1.2.1	7	73.36	91.43
ResNet-50	97.7 MB	MD5	92.0 MB	1.2.1	7	75.81	92.82
ResNet-101	170.4 MB	MD5	159.4 MB	1.2.1	7	77.42	93.61
ResNet-152	230.3 MB	MD5	216.0 MB	1.2.1	7	78.20	94.21

Custom Vision Service: customvision.ai

1. Upload photos and label

2. Train

3. Download ONNX model!



Convert models

- **Tensorflow:** onnx/tensorflow-onnx
- **Keras:** onnx/keras-onnx
- **Scikit-learn:** onnx/sklearn-onnx
- **CoreML:** onnx/onnxmltools
- **LightGBM:** onnx/onnxmltools
- **LibSVM:** onnx/onnxmltools
- **XGBoost:** onnx/onnxmltools
- **SparkML** (alpha): onnx/onnxmltools

Native export

- **Pytorch**
- **CNTK**

Convert models: Examples

```
from keras.models import load_model
import keras2onnx
import onnx

keras_model = load_model("model.h5")

onnx_model = keras2onnx.convert_keras(keras_model,
keras_model.name)

onnx.save_model(onnx_model, 'model.onnx')
```

```
python -m tf2onnx.convert
        --input frozen_model.pb
        --inputs input_batch:0,
lengths:0
        --outputs top_k:1
        --fold_const
        --opset 8
        --output deepcc.onnx
```

Chainer

```
serializers

serializers.load_npz("my.model", model)

sample_input = np.zeros((1, 3, 224, 224), dtype=np.float32)
chainer.config.train = False

onnx_chainer.export(model, sample_input, filename="my.onnx")
```

PyTorch

```
import torch
import torch.onnx

model = torch.load("model.pt")

sample_input = torch.randn(1, 3, 224,

torch.onnx.export(model, sample_input, "model.onnx")
```

ONNX & ONNX Runtime - Community Projects

Get Involved

Discuss

Participate in discussions for advancing the ONNX spec.

gitter.im/onnx

Contribute

Make an impact by contributing feedback, ideas, and code.

github.com/onnx

github.com/microsoft/onnxruntime

Q&A