ONNX

Converter SIG Updates

Kevin Chen (NVIDIA)
Thiago Crepaldi (Microsoft)

Agenda

- Frontend Converters
  - PyTorch
  - TensorFlow
  - SKLearn
  - ONNX Script
- Backend Converters
  - ONNX Runtime
  - TensorRT
- Roadmap
- Get Involved

PyTorch → ONNX

Latest release: PyTorch 2.0.1 + opset 18

Preview: PyTorch nightly features the new TorchDynamo-basedONNX exporter that will ship as part of PyTorch 2.1 (October, 2023):

- New API: `torch.onnx.dynamo_export`
- All Python, no C++
- Built around ONNX Script
  - Aim to implement all of Core ATen IR and Prims IR
- Fake tensor support introduced into TorchDynamo
  - Critical for LLMs, fast exporting, and no OOMs

PyTorch Dynamo → ONNX

Preview!

```
import torch
from transformers import GPT2Tokenizer, GPT2Model

tokenizer = GPT2Tokenizer.from_pretrained("gpt2-medium")
model = GPT2Model.from_pretrained("gpt2-medium")
text = "Replace me by any text you'd like."
encoded_input = tokenizer(text, return_tensors="pt")


model(**encoded_input)


torch.onnx.dynamo_export(
   model, **encoded_input
).save("gpt2_medium.onnx")
```

The Dynamo Exporter now just needs `*args` and `**kwargs` specified exactly as they would be to run the model.

ONNX Script → ONNX

Preview!

ONNX Script is a new open-source library for directly authoring ONNX models in Python with a focus on clean, idiomatic Python syntax and composability through ONNX-native functions.

Critically, it is also the foundation upon which we are building the new PyTorch ONNX exporter to support TorchDynamo – the future of PyTorch.

Example: $\mathrm{GELU}(x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}(x/\sqrt{2})\right]$

```python
import math
from onnxscript import script, opset18 as op, FLOAT

M_SQRT1_2 = math.sqrt(0.5)

@script()
def gelu(X: FLOAT[...]):
    phiX = 0.5 * (op.Erf(M_SQRT1_2 * X) + 1.0)
    return X * phiX
```

TensorFlow → ONNX

Latest release: tf2onnx 1.14.0 + opset 18, up to TF 2.11

- Default opset export version updated to 15
- Added `--outputs_as_nchw` export option to automatically insert transposes at network outputs
- Added support for: tf.math.cumprod, BatchMatMulV3
- Fixes for loop exports into ONNX
- https://github.com/onnx/tensorflow-onnx/releases

SKLearn → ONNX

Latest release: sklearn-onnx 1.14.1 + opset 15

- Added support for sklearn==1.12
- Added support for: OneVsOne, QuadraticDiscriminantAnalysis, GammaRegressor, _ConstantPredictor
- Various bug fixes
- https://github.com/onnx/sklearn-onnx/releases

ONNX → TensorRT

Latest release: onnx-tensorrt 8.6 + opset 17

- New op support: GroupNormalization, LayerNormalization, IsInf
- Improved dynamic shape support for ReverseSequence, Trilu, and TopK
- Updated casting semantics
- Added TensorRT layer annotation with ONNX metadata
- https://github.com/onnx/onnx-tensorrt/releases/

ONNX → ONNX Runtime

Latest release: v1.15.1

- ONNX 1.14.0 support (opset 19)
- Introduces training on edge devices (Desktop and Android)
- Support CUDA 11.4, 11.8, 12.x and Python 3.8 - 3.11
- Improved performance over stock PyTorch for
  - ViT, BEIT and SwinV2 up to 44% speedup with ORT + DeepSpeed
  - Popular HuggingFace models speedup 4% to 15%
  - StableDiffusion, GPT, T5, whisper models also had speedup
- Support different hardware
  - Support Web, Mobile, OpenVino, DirectML, TensorRT, ROCm, and AzureEP
  - Two new execution providers added: JS EP and QNN EP

Roadmap

Minimal opset support - WIP

- Formal proposal to deprecate opset < 9 in progress

Improved community-driven tooling - WIP

- Fuzz-testing, shape / type inference, quantization tools, constant-folding

Get Involved!

- Feedback? Join us on Slack in the #onnx-converters channel
- Subscribe to ONNX Converters SIG mailing list
- Open up issues and partake in discussions on Github