Intel® Neural Compressor

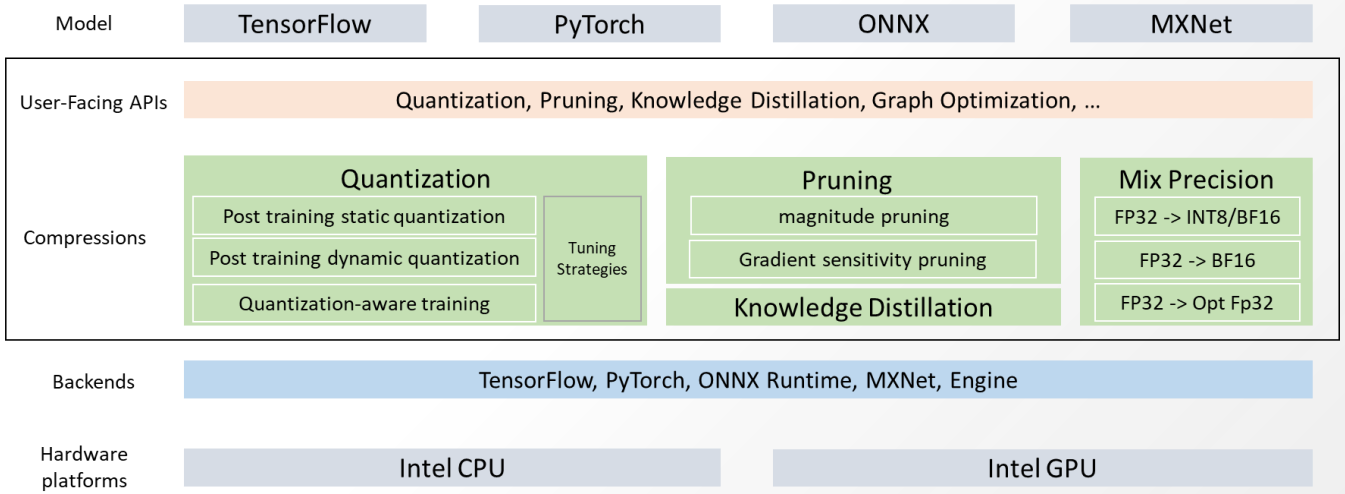
A Scalable Quantization Tool for ONNX Models

Intel® Neural Compressor

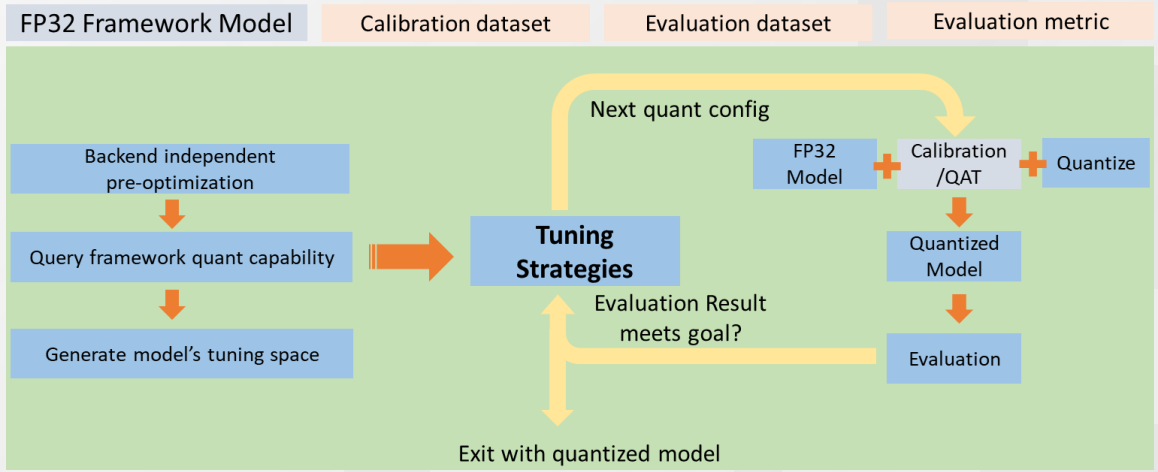
Intel® Neural Compressor is an open-source Python library running on Intel CPUs and GPUs, which delivers unified interfaces across multiple deep learning frameworks for popular network compression technologies, such as quantization, sparsity/pruning and knowledge distillation.

- Verified HWs: Xeon (SKX/CLX/CPX/ICX/SPR)

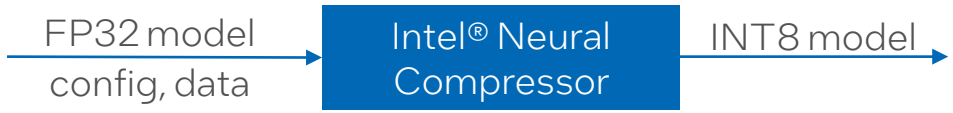
Intel® Neural Compressor Architecture



Auto-tuning Flow



Deploy ONNX model Quantization Rapidly



Based on built-in components of Intel® Neural Compressor, user can quantize a model with config and just 5 lines of code.

Typical Built-in Dataset & Transform & Metric

- Dataset:

ImageFolder, ImagenetRaw, COCORaw, GLUE, ...

- Transform:

Resize, CenterCrop, Normalize, ...

- Metric:

topk, mAP, GLUE, ...

- Config

```
model:
  name: resnet50_v1_5
  framework: onnxrt_qlinearops
quantization:
  approach: post_training_static_quant
  calibration:
    dataloader:
      dataset:
        …
      transform:
        …
evaluation:
  accuracy:
    metric:
      …
    dataloader:
      dataset:
        …
      transform:
        …
```

- Launch code

```
from neural_compressor.experimental \
    import Quantization, common
quantize = Quantization(args.config)
quantize.model = common.Model(model)
q_model = quantize()
q_model.save(args.output_model)
```

Contribution to ONNX Model Zoo

- Use Intel® Neural Compressor to generate quantized models and upstream to ONNX Model Zoo .

Model	Version	Model Size(MB)		Accuracy(%)		Accuracy Drop(%): (INT8-FP32)/FP32	Performance Improvement
		FP32	INT8	FP32	INT8		
Resnet50_v1	1.9.0 (opset11+)	97.8	24.6	74.97	74.83	0.19	1.85x
VGG16		527.8	132.0	72.38	72.37	0.01	1.54x
Shufflenetv2		8.79	2.28	66.35	66.15	0.30	1.72x
BERT-MRPC		417.72	106.76	86.03	85.54	0.57	2.45x
BERT-Squad		415.66	118.80	80.67	80.44	0.29	1.81x
RoBERTa-MRPC		475.55	126.01	88.73	89.46	0.82	2.43x
Distilbert-MRPC		255.44	65.74	84.56	84.56	0.00	2.80x

*INT8 Resnet50 is the first quantized model for ONNX model zoo.

*Resnet50, VGG16, Shufflenetv2 has been upstreamed to ONNX model zoo, other models are working in progress.

*The performance depends on the test hardware. Performance data here is collected with Intel® Xeon® Platinum 8280 Processor, 1s 4c per instance, CentOS Linux 8.3.

Contribution Plan

All enabled FP32 models in ONNX model zoo would have corresponding quantized modes through Intel® Neural Compressor.

- First Stage
  - Image Classification & Domain-based Image Classification models
  - Object Detection & Image Segmentation models
- Second Stage
  - Machine Comprehension models
  - Speech & Audio Processing models
  - Image Manipulation models
  - Body, Face & Gesture Analysis models


intel