ONNX Workshop Mar 24 2021

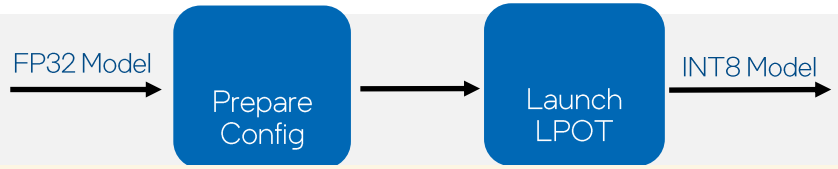
Quantization support for ONNX using Intel® Low Precision Optimization Tool (LPOT)

Haihao Shen and Saurabh Tangri

Intel® Low Precision Optimization Tool

Intel® Low Precision Optimization Tool (LPOT) is helping Intel customers rapidly deploy low-precision inference solution for popular deep learning (DL) frameworks on CPU and GPU.

- Mixed precisions: INT8, BF16*, and FP32
- Verified HWs: Xeon (SKX/CLX/CPX/ICX/SPR), Xe



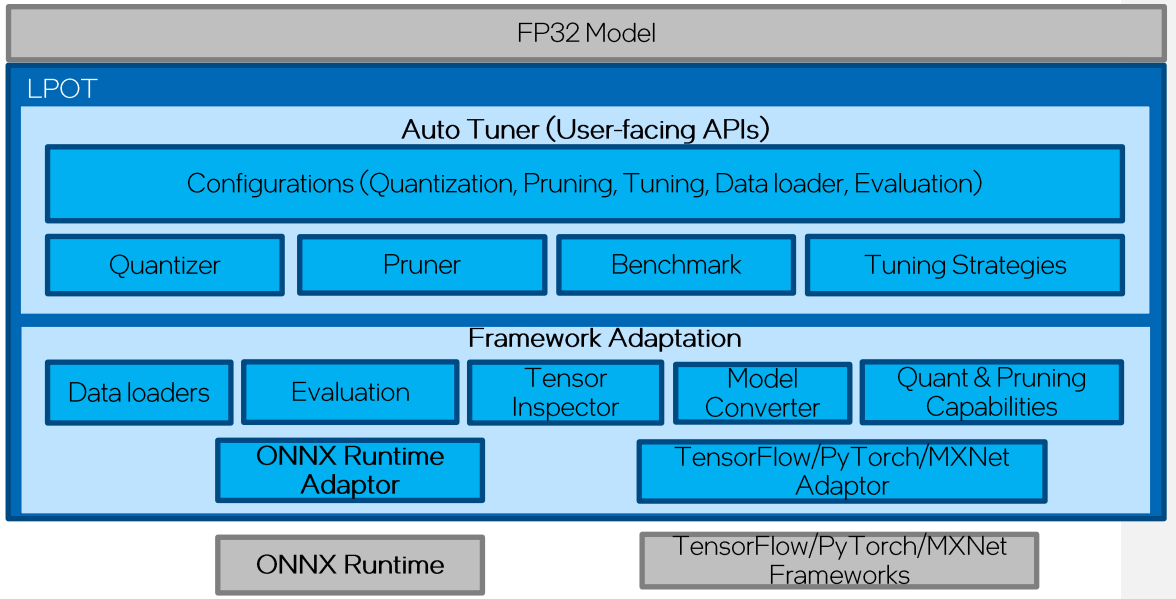
Template-based config

```
model:
  name: resnet50_v1_5
  framework: onnxrt_qlinearops
quantization:
  approach: post_training_static_quant
  calibration:
    dataloader:
      dataset:
        ImageFolder:
          root: /path/to/calibration/dataset/
      transform:
        ResizeCropImagenet:
          height: 224
          width: 224
          mean_value: [0.485, 0.456, 0.406]
```

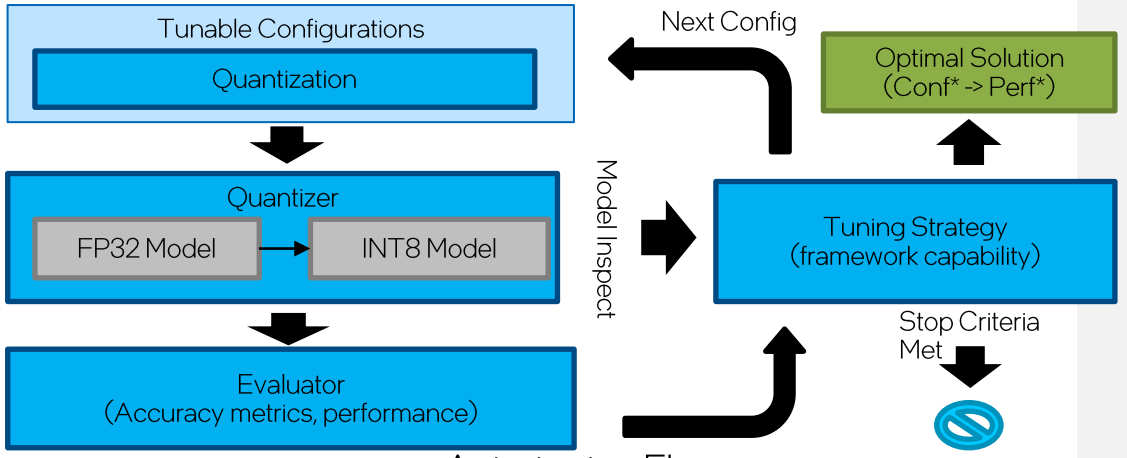
Launch code

```
from lpot.quantization import Quantization
quantize = Quantization(args.config)
q_model = quantize(model)
q_model.save(args.output_model)
```

*: depend on kernel readiness on frameworks

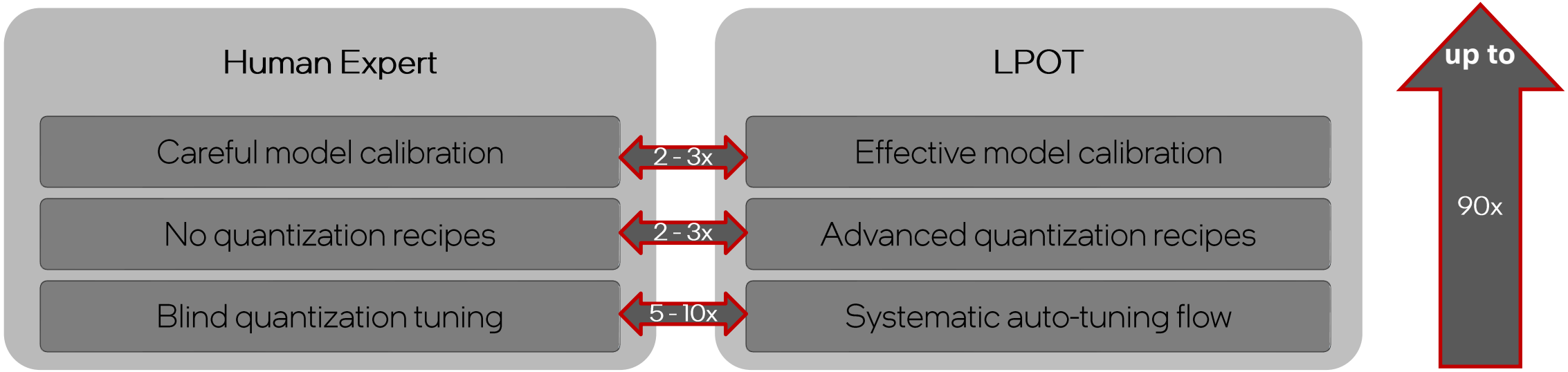


LPOT Architecture



Auto-tuning Flow

Operationalize Quantized models from Days to Mins



Framework	Version	Model	Dataset	Tuning Time (s)	Accuracy		
					INT8	FP32	Relative Loss: (INT8-FP32)/FP32
ONNX RT	1.6.0 (opset11+)	resnet50_v1_5	ImageNet	1361	73.60%	74.00%	-0.54%
ONNX RT		vgg16	ImageNet	2383	68.86%	69.44%	-0.84%
ONNX RT		bert_base_mrpc	MRPC	30	85.29%	86.03%	-0.85%
ONNX RT		MobileBERT	MRPC	44	0.8603	0.8627	-0.28%
ONNX RT		RoBERTa	MRPC	76	0.8873	0.8946	-0.82%
ONNX RT		DistilBERT	MRPC	42	0.8505	0.8456	0.58%

Quantization support: 1) two quantized op categories: QLinearOps & IntegerOps; 2) two quantization approaches: static & dynamic

Releases, Collaborations, and Plans

- Releases
  - v1.2 (WW11'21): ONNX RT v1.6, operator-wise quantization tuning
  - v1.3 (WW16'21): ONNX RT v1.7, new quantized operators
  - v1.4 (WW22'21): Python optimizer tool integration
- Community Collaboration
  - Welcome the contributions from community
  - Submit a pull request or file an issue by following contribution guidelines
- Plans
  - ONNX model zoo quantization support
  - Formal release via docker distribution; nightly-built binary release via pip package


intel