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

```
model: Template-based config
  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]
```

* depend on kernel readiness on frameworks
Operationalize Quantized models from Days to Mins

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

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