Intel® Neural Compressor

A Scalable Quantization Tool for ONNX Models
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)
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, ...

```
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)
```

```
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**
```python
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Version</th>
<th>Model Size(MB)</th>
<th>Accuracy</th>
<th>Accuracy Drop</th>
<th>Performance Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FP32</td>
<td>INT8</td>
<td>FP32</td>
<td>INT8</td>
</tr>
<tr>
<td>Resnet50_v1</td>
<td>1.9.0 (opset11+)</td>
<td>97.8</td>
<td>24.6</td>
<td>74.97</td>
<td>74.83</td>
</tr>
<tr>
<td>VGG16</td>
<td></td>
<td>527.8</td>
<td>132.0</td>
<td>72.38</td>
<td>72.37</td>
</tr>
<tr>
<td>Shufflenetv2</td>
<td>1.9.0 (opset11+)</td>
<td>8.79</td>
<td>2.28</td>
<td>66.35</td>
<td>66.15</td>
</tr>
<tr>
<td>BERT-MRPC</td>
<td></td>
<td>417.72</td>
<td>106.76</td>
<td>86.03</td>
<td>85.54</td>
</tr>
<tr>
<td>BERT-Squad</td>
<td></td>
<td>415.66</td>
<td>118.80</td>
<td>80.67</td>
<td>80.44</td>
</tr>
<tr>
<td>RoBERTa-MRPC</td>
<td>1.9.0 (opset11+)</td>
<td>475.55</td>
<td>126.01</td>
<td>88.73</td>
<td>89.46</td>
</tr>
<tr>
<td>Distilbert-MRPC</td>
<td></td>
<td>255.44</td>
<td>65.74</td>
<td>84.56</td>
<td>84.56</td>
</tr>
</tbody>
</table>

*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