Quantization Aware Training Project for Computer Vision Models

⚠ Disclaimer: All datasets hyperlinked from this page are not owned or distributed by Google. The dataset is made available by third parties. Please review the terms and conditions made available by the third parties before using the data.

Overview

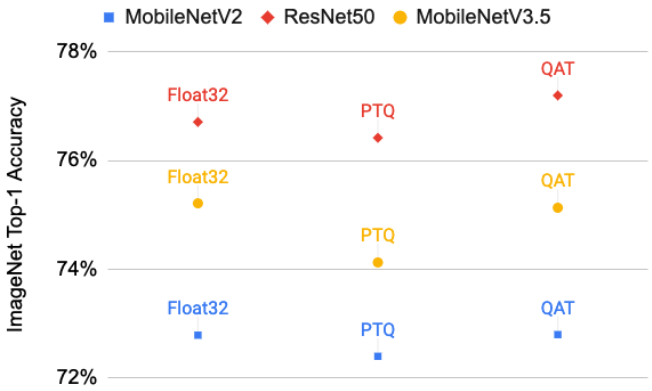
This project includes quantization aware training code for Computer Vision models. These are examples to show how to apply the Model Optimization Toolkit's quantization aware training API.

Note: Currently, we support a limited number of ML tasks & models (e.g., image classification and semantic segmentation) We will keep adding support for other ML tasks and models in the next releases.

How to train a model

```
EXPERIMENT=xxx  # Change this for your run, for example, 'mobilenet_imagenet_qat'
CONFIG_FILE=xxx  # Change this for your run, for example, path of
imagenet_mobilenetv2_qat_gpu.yaml
MODEL_DIR=xxx  #  Change this for your run, for example, /tmp/model_dir
$ python3 train.py \
--experiment=${EXPERIMENT} \
--config_file=${CONFIG_FILE} \
--model_dir=${MODEL_DIR} \
--mode=train_and_eval
```

Image Classification



Comparison of Imagenet top-1 accuracy for the classification models

Note: The Top-1 model accuracy is measured on the validation set of ImageNet.

Pre-trained Models

Model	Resolution	Top-1 Accuracy (FP32)	Top-1 Accuracy (Int8/PTQ)	Top-1 Accuracy (Int8/QAT)	Config	Download
MobileNetV2	224x224	72.782%	72.392%	72.792%	config	TFLite(Int8/QAT)
ResNet50	224x224	76.710%	76.420%	77.200%	config	TFLite(Int8/QAT)
MobileNetV3.5 MultiAVG	224x224	75.212%	74.122%	75.130%	config	TFLite(Int8/QAT)

Semantic Segmentation

Model is pretrained using COCO train set. Two datasets, Pascal VOC segmentation dataset and Cityscapes dataset (only for DeepLab v3+), are used to train and evaluate models. Model accuracy is measured on full Pascal VOC segmentation validation set.

Pre-trained Models

model | resolution | mIoU | mIoU (FP32) | mIoU (FP16) | mIoU (INT8) | mIoU (QAT INT8) | download (tflite) :------------------------- | :--------: | ----: | ----------: | ----------: | ----------: | --------------: | --: | ---: | --: | --: | ----------------: MobileNet v2 + DeepLab v3 | 512x512 | 75.27 | 75.30 | 75.32 | 73.95 | 74.68 | FP32 | FP16 | INT8 | QAT INT8 MobileNet v2 + DeepLab v3+ | 1024x2048 | 73.82 | 73.84 | 73.65 | 72.33 | 73.49 | FP32 | FP16 | INT8 | QAT INT8