Run Quantized Model

**Step 1**

The first step is to obtain the fractional bits of inputs, outputs and weights of every layer based on the original model and test dataset.

The information of all the fractional bits are stored in the form of { “layername” :{ “weights”: value, “biases”:value} } as a numpy dictionary. These data are stored in tf-reference-model\lib\quantized\_model, where fcn8\_seg, ssd and rcnn correspond to the three 02\_task\_specific\_benchmarks models (01\_semseg, 02\_objdet and 03\_actrec), respectively.

The information of fractional bits and saved quantized weights (multiplied with fractional bits) are used to rebuild the model, which are used to code the fcn8net.py, ssdnet.py and rcnnnet.py under tf-reference-model\lib\networks. The method to rebuild the quantized model please refer to the documents of refmodel.

**Step 2**

The rebuilt model is spliced by the quantized parts and the unclear parts from the original model. The quantized parts are built based on the layer definition of network.py in fcn8net.py, ssdnet.py, and rcnnnet.py. Since there are some operations without parameters that cannot be manually re-implemented precisely, we crop this sub-model graph from the original model, then fuse and slice the quantized parts with the sub-model using fusion\_graph function in fcn8net.py, ssdnet.py, and rcnnnet.py to obtain the whole model.

Before the whole model run the test data, it first calls the load function from network.py to assign the quantized weights to its corresponding weights in the model.

The original model, cropped sub-model, fused whole model (all\_quantized\_model.pb), the crop scripts, the results from Bosch, the quantized and original results from ref model are all located at tf-reference-model\lib\tensorflow\_model.

**Step 3**

The run scripts of the three models are all located at tf-reference-model\lib\evaluation. The preprocessing method is the same as Bosch provided. The only difference is we are running quantized model. All the model location can refer to .sh files.

In /private/liusongwei/

Results

|  |  |  |  |
| --- | --- | --- | --- |
|  | 01\_semseg | 02\_objdet | 03\_actrec |
| Bosch\_results  result.txt | OPA：0.860 | LAMR: 0.529 | OPA: 0.868 |
| OMA: 0.368 |  | OMA：0.641 |
| Original\_results  (Ref Model)  test\_result.txt | OPA：0.854 | LAMR: 0.529 | OPA: 0.868 |
| OMA: 0.340 |  | OMA: 0.641 |
| Quantized\_results  (Ref Model)  quantized\_result.txt | OPA：0.831 | LAMR: | OPA: 0.862 |
| OMA: 0.321 |  | OMA: 0.655 |

The detailed results are stored in .txt files inside lib\tensorflow\_model\\*\reference\_result.