模型量化

总体说明：

1：第一步由原模型和测试数据，得到模型每一层的输入，输出，权重，偏置小数位信息。

所有的小数位信息以{ “layername” :{ “weights”: value, “biases”:value} } 字典的形式存放，存放于tf-reference-model\lib\quantized\_model 文件夹中，其中fcn8\_seg, ssd, rcnn 分别对应第1，2，3个任务模型。

小数位信息和保存好的量化权重（已经乘上了小数位）用于重构模型，也就是编写tf-reference-model\lib\networks文件夹下面的 fcn8net.py, ssdnet.py, rcnnnet.py , 模型重构的方式参考refmodel说明文件。

2：在重构模型的时候使用了模型拼接的方法，在fcn8net.py, ssdnet.py, rcnnnet.py 构建了模型能够被量化的部分（使用network.py里面的层定义重构网络），然后由于原模型当中存在一些无参数操作无法准确地手动复现，因此从原模型当中裁剪下这部分子图，然后在fcn8net.py, ssdnet.py, rcnnnet.py这些文件的fusion\_graph 函数中将重构部门和裁剪的原模型的子图进行融合拼接得到完整的模型。

完整的模型在运行测试数据之前，先调用network里面的load方法，将第一步得到各个模型的量化后的权重赋值给模型中对应的权重。

原模型，裁剪子模型，拼接得到的完整量化模型，裁剪脚本，还有各个模型的Bosch提供的参考结果，本平台的原模型的实测结果，量化模型的实测结果，均位于tf-reference-model\lib\tensorflow\_model 中。

3：三个模型的测试脚本位于tf-reference-model\lib\evaluation，所有的数据的预处理方式跟Bosch所提供的测试脚本相同，只是这里跑的是量化模型。 所有的数据集位置可参考对应运行脚本.sh文件.

位于/private/liusongwei/ 中。

测试结果：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 01\_semseg | 02\_objdet | 03\_actrec |
| Bosch\_results | OPA：0.860 | LAMR: 0.529 | OPA: 0.868 |
| OMA: 0.368 |  | OMA：0.641 |
| True\_results | OPA：0.854 | LAMR: 0.529 | OPA: 0.868 |
| OMA: 0.340 |  | OMA: 0.641 |
| Quantized\_results | OPA：0.831 | LAMR: 0.526 | OPA: 0.862 |
| OMA: 0.321 |  | OMA: 0.655 |

详情见相应的存放结果的.txt文件