Tengine Quant Tool User Manual

文档版本 1.4

发布日期 2021/06/25

OPEN AI LAB

变更记录

日期	版本	说明	作者
2020-12-04	1.0	初版	Tang Qi
2021-04-02	1.1	更新 2.1 章节量化工具指令参数说明	Tang Qi
2021-06-22	1.2	更新 2.1 章节量化工具指令参数说明， 新增 KL 量化算法，和 letterbox 参数	Tang Qi
2021-06-25	1.4	部分参数说明修正，增加小提示	Chen Honghao

目录

1 Tengine Quant Tool 简介

Tengine Quant Tool 是针对 Tengine 进行的模型量化工具，支持将 Tengine 的 FP32 模型量化、压缩成 UINT8 模型。

2 Tengine Quant Tool 使用说明

2.1 量化工具指令参数说明:

参数设置如下:

参数名	说明
-h	量化工具参数说明
-m	输入的 Float32 格式的 tmfile 文件
-i	输入的量化校准使用的数据集文件夹路径
-o	输出的 UInt8 格式的 tmfile 文件
-a	量化算法选择（0：MIN MAX，1：KL）
-g	模型 input dims 输入（default is 3,224,224）
-w	模型 mean 输入（default is 104,117,123）
-s	模型 scale 输入（default is 1,1,1）
-b	图片预处理RGB或BGR（0：BGR，1：RGB，default is 1）
-c	图片预处理是否进行 Center Crop 处理（0：OFF，1：ON，default is 0）
-y	预处理采用 Letter Box 处理（[letterbox_rows, letterbox_cols], default is [0,0]）

-t	运行线程数设置（default is 4）

2.2 量化功能

以基于 MobileNet v1 模型为例，分别说明各种参数的使用方式。输入模型文件是 mobilenet_fp32.tmfile。

2.2.1 量化校准策略

此版本支持两种量化校准策略。

2.2.1.1 min-max 量化校准策略

对模型中每一层输出 activation、weight 数据进行最大最小值统计，计算生成 scale 和 zero point 数值：

```
./quant_tool_uint8 -m ./mobilenet_fp32.tmfile -i ./dataset -o ./mobilenet_uint8.tmfile -g 3,224,224 -w 104.007,116.669,122.679 -s 0.017,0.017,0.017
```

2.2.1.2 kl 量化校准策略

对模型中每一层输出 activation、weight 数据进行最大最小值统计，计算生成 scale 和 zero point 数值：

```
./quant_tool_uint8 -m ./mobilenet_fp32.tmfile -i ./dataset -o ./mobilenet_uint8.tmfile -g 3,224,224 -w 104.007,116.669,122.679 -s 0.017,0.017,0.017 -a 1
```

2.2.1.3 量化校准建议

当选择 min-max 量化校准策略时，建议选用校准图片 100 张左右。当选择 kl 量化校准策略时，因为需要计算数据分布，建议选用校准图片 300 张以上。校准图片最好具有一定的代表性：符合实际应用场景，且类别不单一。

2.2.2 输出文件说明

- table_minmax.scale：采用 min-max 量化校准策略生成的量化校准表文件；

- table_kl.scale：采用 min-max 量化校准策略生成的量化校准表文件；
- xxx.tmfile：UInt8 格式的 tmfile。

3 技术支持

如有技术问题请联系：support-tengine@openailab.com