Tengine Quant Tool User Manual

文档版本 1.0

发布日期 2020/12/04

OPEN AI LAB

变更记录

日期	版本	说明	作者
2020-12-04	1.0	初版	Tang Qi

目录

1 Tengine Quant Tool 简介

Tengine Quant Tool 是 Tengine 推出的一款 x86 Linux 模型量化工具。

主要支持以下功能：

- UInt8 量化功能：将 Caffe / TensorFlow / MXNet / ONNX / TFLite / Tengine 等框架的 Float32 模型量化为 Tengine 的 UINT8 模型。量化功能分为两种模式:
  - EXTERNAL，加载外部已有量化表，获取量化参数，对模型进行量化；
  - INTERNAL，通过内部量化模块，生成量化参数，对模型进行量化。

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	工具显示说明
-f	框架类别，支持框架为：Caffe/TensorFlow/MXNet/ONNX/TFLite/Darknet
-p	模型参数文件输入（需包含文件路径）
-m	模型文件输入（需包含文件路径）
-s	如果是外部量化模式，需要输入量化参数表（需包含量化文件路径）
-o	输出文件名（需包含输出文件路径）
-t	转化类别，现提供 UINT8 端到端模型转化
-c	量化类别，EXTERNAL外部量化（需要量化参数表），INTERNAL（内部量化）需要载入前端推理框架，现与Tengine适配
-a	量化算法选择：MINMAX，KL
-x	模型均值输入：参考 128，128，128
-y	模型scale输入：参考 55，55，55

-z	模型尺寸输入：224，224，3（height, width, channel)
-i	量化模式图片集路径
-n	量化图片数目

2.2 量化功能

以基于 Caffe 框架的 MobileNetv1 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.prototxt 和 mobilenet.caffemodel。

2.2.1量化精度算法调优

此版本增加两种量化精度算法调用功能

2.2.1.1 MINMAX 调优算法

对模型中 layer 的输出输出进行最大最小值获取进行 scale 与 zero point 计算

```
./convert_tool -f caffe -m mobilenet.caffemodel -p mobilenet.prototxt -o mobilenet_minmax.tmfile -a MINMAX -i /tmp/tiny_voc/ -x 127.5,127.5,127.5 -y 57.8,57.6,57.5 -z 224,224,3 -c INTERNAL -t UINT8 -n 100
```

通过对 -a 进行算法选取，算法选项为 -a MINMAX

2.2.1.2 KL 调优算法

Kullback-Leibler Divergence 相对熵算法，根据模型输入输出的范围分布，截取合理的最大最小值进行 scale 与 zero point 计算

```
./convert_tool -f caffe -m mobilenet.caffemodel -p mobilenet.prototxt -o mobilenet_kl.tmfile -a KL -i /tmp/tiny_voc/ -x 127.5,127.5,127.5 -y 57.8,57.6,57.5 -z 224,224,3 -c INTERNAL -t UINT8 -n 100
```

算法选项为 -a KL，因为 KL 算法需根据图片来计算数据分布，所以图片选取数量需根据具体模型进行实验所得，再通过 -n 来指定图片数量

2.2.2 输出文件说明

经过量化工具转换，其生成的文件包含三个类型：

1. .tmfile 结尾：原 FP32 模型
2. .tmfilefinetunescale 结尾：量化模型最终每一层量化表
3. .tmfileoutscale 结尾：模型经过量化算法后生成的量化表
4. _UINT8.tmfile: UINT8 量化模型
5. _FP32.tmfile: FP32 模型附带 scale 与 zero point 参数

3 技术支持

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