

Tengine Quant Tool

User Manual

文档版本 1.0

发布日期 2020/12/04

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 3](#_Toc60737060)

[2 Tengine Quant Tool 使用说明 3](#_Toc60737061)

[2.1 量化工具指令参数说明： 3](#_Toc60737062)

[2.2 量化功能 4](#_Toc60737063)

[2.2.1 量化精度算法调优 4](#_Toc60737064)

[2.2.1.1 MINMAX调优算法 4](#_Toc60737065)

[2.2.1.2 KL调优算法 4](#_Toc60737066)

[2.2.2 输出文件说明 4](#_Toc60737067)

[3 技术支持 5](#_Toc60737068)

# 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。

# Tengine Quant Tool 使用说明

## 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 |
| -q | 工具类别： QUANT(量化功能)，OPERATOR(显示框架支持算子列表) |
| -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 |

## 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。

### 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。

#### 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。



通过对 -a 进行算法选取，算法选项为 -a MINMAX。

#### 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。



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。

### 输出文件说明

经过量化工具转换，其生成的文件包含五个类型：

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。

# 技术支持

如有技术问题请联系：[support-tengine@openailab.com](file:///C:\Users\Lenovo\AppData\Roaming\Microsoft\Word\support-tengine@openailab.com)