

Tengine Quant Tool

User Manual

Version 1.0

Release Date 12/04/2020

Change Logs

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Description | Author |
| 2020-12-04 | 1.0 | first edition | Qi Tang |

Contents

[1 Brief 3](#_Toc60756881)

[2 Instruction 3](#_Toc60756882)

[2.1 Parameter description of quantization tool instruction: 3](#_Toc60756883)

[2.2 How to use this tool 4](#_Toc60756884)

[2.2.1 The algorithm of tuning quantization accuracy 4](#_Toc60756885)

[2.2.1.1 MINMAX tuning algorithm 4](#_Toc60756886)

[2.2.1.2 KL tuning algorithm 5](#_Toc60756887)

[2.2.2 The description of output file 5](#_Toc60756888)

[3 Technical support 5](#_Toc60756889)

# Brief

Tengine Quant Tool is an x86 Linux model quantification tool development by Tengine team that can run on Ubuntu 18.04.

Mainly supports the following functions:

* Quantification function of UInt8: Quantify Float32 model of Caffe/TensorFlow/Mxnet/ONNX/TFLite/Tengine into UINT8 model of Tengine. The quantization function is divided into two modes:
  + EXTERNAL, loading the external existing quantization table, obtaining quantization parameters, and quantizing the model;
  + INTERNAL, through the internal quantization module, generates quantization parameters and quantizes the model.

# Instruction

## Parameter description of quantization tool instruction:

|  |  |
| --- | --- |
| Params | Description |
| -h | Tool display description |
| -q | Tool category: QUANT (quantization function), OPERATOR (display frame support operator list) |
| -f | Framework category, supported by Caffe/TensorFlow/Mxnet/ONNX/TFLite/Darknet |
| -p | Model parameter file input (including file path) |
| -m | Model file input (including file path) |
| -s | In case of external quantization mode, you need to input quantization parameter table (including quantization file path) |
| -o | Output file name (including output file path) |
| -t | Transformation category, now provides UINT8 end-to-end model transformation |
| -c | Quantization category, EXTERNAL external quantization (requires quantization parameter table), and INTERNAL (internal quantization) need to be loaded into the front-end reasoning framework, which is now adapted to Tengine |
| -a | Selection of quantization algorithm: MINMAX, KL |
| -x | Model mean input (mean value, default is 128,128,128) |
| -y | Model normalize input (scale value, default is 55,55,55) |
| -z | Model size input: 224,224,3 (height, width, channel) |
| -i | Path of calibration picture set |
| -n | The num of calibration images |

* 1. How to use this tool

Taking MobileNetv1 model based on Caffe framework as an example, the usage modes of various parameters are explained respectively. The input model files are mobilenet.prototxt and mobilenet.caffemodel.

### The algorithm of tuning quantization accuracy

In this version, two quantization precision algorithm calling functions are added.

#### MINMAX tuning algorithm

Get the maximum and minimum values of the output of layer in the model, and calculate scale and zero point.

Through the algorithm selection of -a, the algorithm option is -a MINMAX.

#### KL tuning algorithm

The relative entropy algorithm of Kullback-Leibler Divergence, according to the range distribution of input and output of the model, intercepts the reasonable maximum and minimum values for scale and zero point calculation.

The algorithm option is -a KL, because KL algorithm needs to calculate the data distribution according to pictures, so the number of pictures should be obtained by experiments according to specific models, and then the number of pictures should be specified by -n.

### The description of output file

After being transformed by the quantization tool, the generated files contain five types:

1. .tmfile: The original Float32 tmfile;
2. .tmfilefinetunescale: The final calibration quantization table;
3. .tmfileoutscale: The calibration quantization table;
4. \_UINT8.tmfile: The uint8 tmfile;
5. \_FP32.tmfile: The Float32 tmfile with the scale and zero point params.

# Technical support

Any other questions, please contact：[support-tengine@openailab.com](file:///C:\Users\Lenovo\AppData\Roaming\Microsoft\Word\support-tengine@openailab.com)