Tengine Quant Tool
User Manual

Version 1.4
Release Date 06/25/2021

OPEN AI LAB

Development Logfile

Date,	Version	explain	Author
2020-12-04	1.0	First edition	Tang Qi
2021-04-02	1.1	Update the instruction parameter description of quantization tool in chapter 2.1	Tang Qi
2021-06-22	1.2	Updated instruction parameter description of quantization tool in chapter 2.1, added KL quantization algorithm, and letterbox parameter	Tang Qi
2021-06-25	1.4	Some parameter descriptions are corrected, and small tips are added	Chen Honghao

OPEN AI LAB

Contents

OPEN AI LAB

1 Brief

Tengine Quant Tool is a model quantization tool for Tengine, which supports the quantization and compression of Tengine's FP32 model into a UINT8 model.

2 Instruction

2.1 Parameter description of quantization tool instruction:

Parameter settings are as follows:

Params	Description
-h	Parameter description of quantization tool
-m	Tmfile file in Float32 format entered
-i	The path of the dataset folder used by the quantitative calibration entered
-o	Output tmfile file in UInt8 format
-a	Selection of quantization algorithm (0: minmax, 1: KL)
-g	The model input dims (default is 3,224,224)
-w	The model input means(default is 104,117,123)
-s	The model input std (default is 1,1,1)
-b	Image preprocessing RGB or BGR (0: BGR, 1: RGB, default is 1)
-c	Is the picture preprocessed by Center Crop (0: NO, 1: Yes, default is 0)
-y	The Letter Box is used for preprocessing ([letterbox _ rows, letterbox _ cols], default is [0,0])
-t	Set the number of running threads (default is 4)

2.2 How to use this tool

Taking the model based on MobileNet v1 as an example, the usage modes of various parameters are explained respectively. The input model file is mobilenet_fp32.tmfile.

2.2.1 Quantitative calibration strategy

This version supports two quantitative calibration strategies.

2.2.1.1 Min-max quantitative calibration strategy

Make statistics on the maximum and minimum values of activation and weight data output from each layer in the model, and calculate and generate scale and zero point values:

```
./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 quantization calibration strategy

Make statistics on the maximum and minimum values of activation and weight data output from each layer in the model, and calculate and generate scale and zero point values:

```
./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 Quantitative calibration recommendations

When choosing the min-max quantitative calibration strategy, it is recommended to choose about 100 calibration pictures. When choosing kl quantization calibration strategy, it is recommended to choose more than 300 calibration pictures because of the need to calculate the data distribution. Calibration pictures should be representative to a certain extent: they conform to the actual application scenarios, and the categories are not single.

2.2.2 Output file description

- Table_minmax.scale: the quantitative calibration table file generated by the min-max quantitative calibration strategy;
- Table_kl.scale: the quantitative calibration table file generated by the min-max quantitative calibration strategy;
- Xxx.tmfile: tmfile in uint8 format.

3 Technical support

If you have any technical questions, please contact: support-tengine@openailab.comsupport-tengine@openailab.com