



Tengine Quant Tool Technical Spec

文档版本 1.1

发布日期 2021-04-02

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 2.4 UInt8 量化模型支持 章节 更新 附录 1 Tengine Quant Tool 算子支持列表	Tang Qi

目录

1	产品介绍	3
1.1	背景与目的	3
1.2	产品特点	3
2	支持范围	3
2.1	硬件支持	3
2.2	操作系统支持	3
2.3	算子支持	3
2.3.1	Tengine 算子支持	3
2.4	UINT8 量化模型支持	3
2.4.1	分类模型	4
2.4.2	检测模型	4
2.4.3	识别模型	4
附录 1	TENGINE QUANT TOOL 支持算子列表	0

1 产品介绍

1.1 背景与目的

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。

1.2 产品特点

- 1) 此工具支持模型压缩、量化功能。

2 支持范围

2.1 硬件支持

2.2 操作系统支持

- Ubuntu 18.04 以上

2.3 算子支持

2.3.1 Tengine 算子支持

详见附录 1。

2.4 UINT8 量化模型支持

以下模型为已验证过的 UInt8 tmfile，该表会不断更新。

2.4.1 分类模型

MobileNet v1	MobileNet v2	ResNet18	ResNet50
SqueezeNet v1.1	VGG16		

2.4.2 检测模型

YOLOv3	YOLOv5s

2.4.3 识别模型

待更新。

附录 1 Tengine Quant Tool 支持算子列表

Clip	Concat	Const	Convolution	Depthtospace	Dropout
Eltwise	Elu	Fullyconnect	Flatten	Hardswish	Input
Interp	Permute	Pool	Prelu	Relu	Relu6
Reshape	Sigmoid	Slice	Softmax	Spacetodepth	Tanh
Transpose	Upsample	Deconvolution			

OPEN AI LAB