Tengine Quant Tool Technical Spec

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 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，支持将现有多框架的 FP32 模型转换并量化、压缩成 UINT8 模型。

1.2 产品特点

1) 此工具支持模型转化、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。

卷积计算方法包括:

- Direct Convolution
- Winograd Convolution
- Gemm Convolution

2.4 UINT8 量化模型支持

2.4.1 Caffe 模型支持

Shufflenet_v2	**googlenet**	**inception_v3**
mobileface	Mobilenet_v1	mobilenet_ssd
squeeznet	**resnet50**	vgg16
Mobilenet_v3	Mobilenet_v2	

2.4.2 ONNX 模型支持

Shufflenet_v2	**googlenet**	**inception_v3**
mobileface	Mobilenet_v1	mobilenet_ssd
squeeznet	**resnet50**	vgg16
Mobilenet_v3	Mobilenet_v2	

2.4.3 MXNet 模型支持

Retinaface

附录 1 Tengine Quant Tool 支持算子列表

TENGINE	Caffe		MXNet		TensorFlow	TF-lite	ONNX	Darknet
ACCURACY	√							
BATCHNORMALIZATION	BatchNorm		BatchNorm		FusedBatchNorm		√	
		ComposedBN						
RESIZE						RESIZE_NEAREST_NEIGHBOR		
CONCAT	√	√	ConcatV2	CONCATENATION	√	route		
CONST								
CONVOLUTION	√	√	Conv2D	CONV_2D	Conv	convolutional		
		DepthwiseConvolution				DepthwiseConv2dNative	DEPTHWISE_CONV_2D	
	ConvolutionDepthwise							
DECONVOLUTION		√		√		Conv2DBackpropInput		
DETECTIONOUTPUT	√							
DROPOUT		√		Copy		√		√	yolo
ELTWISE	√	_minus_scalar	Add	ADD	Add	shortcut		
				_mul_scalar		Sub	SUB	Sub
		elemwise_add		PROD				
						Rsqrt	RSQRT	
		_div_scalar	RealDiv	DIV	Div			
						Log	LOG	
			Exp	EXP	Exp			
						Pow	POW	
			Sqrt	SQRT				
						Floor	FLOOR	Floor
			Mul	MUL	Mul			
						Minimum		
			AddN					
FLATTEN		√		√		√		√
FULLYCONNECTED	InnerProduct	√	MatMul	FULLY_CONNECTED	MatMul			
								Gemm
INPUT	Data		FIFOQueueV2					
		Input						
LRN	√		√					
NORMALIZE		√						
PERMUTE	√	transpose						

TENGINE		Caffe		MXNet		TensorFlow	TF-lite	ONNX	Darknet
POOLING		√		√		AvgPool	AVERAGE_POOL_2D	AveragePool	
					GlobalAveragePool				
						MaxPool	MAX_POOL_2D	MaxPool	maxpool
PRELU	√	LeakyReLU			PRelu				
PRIORBOX		√							
REGION	√					region			
RELU		√		Activation		Relu		Relu	
		LeakyReLU			LeakyRelu				
RELU6		√		clip		Relu6			
REORG	√					reorg			
RESHAPE		√		√		√	RESHAPE	√	
ROIPOOLING	√								
RPN		√							
SCALE	√								
SLICE		√						√	
SOFTMAX	√	Activation	√	SOFTMAX	√				
		SoftmaxWithLoss							
		SoftmaxOutput							
				SoftmaxActivation					
SPLIT	√		√		√				
DETECTIONPOSTPROCESS							TFLite_Detection_PostProcess		
GEMM									
GENERIC			DecodeWav						
						AudioSpectrogram			
			Mfcc						
LOGISTIC							LOGISTIC		
LSTM		RNN	√						
RNN						√			
TANH	TanH	Activation	√		√				
SIGMOID		√		Activation		√		√	
SQUEEZE				SQUEEZE	√				
PAD						√			
			MirrorPad						
STRIDEDSLICE						√	STRIDED_SLICE		
REDUCTION	√	√	Sum	SUM					
						Mean	MEAN	ReduceMean	
			Asum						
						Sqsum			
			Max						
						Min			
			Prod						
						L2			

TENGINE		Caffe		MXNet	TensorFlow	TF-lite	ONNX	Darknet
			Logsum					
					Logsumexp			
ARGMAX			√					
ARGMIN					√			
TOPKV2			√					
MAXIMUM					√		Max	
MINIMUM			√					
ADDN				add_n				
SWAPAXIS		√						
GRU				RNN	√			
UPSAMPLE	√	UpSampling			upsample			
SHUFFLECHANNEL		√						
RESIZE	√		ResizeNearestNeighbor					
					ResizeBilinear			
SPACETOBATCHND			√					
BATCHTOSPACEND					√			
CROP	√	√						
PSROIPOOLING				_contrib_PSROIPooling				
ROIALIGN		_contrib_ROIAlign						
EXPANDDIMS					ExpandDims			
UNARY			√					
				abs	Abs			
		neg	Neg					
				ceil	Ceil			
		floor	Floor					
				sin	Sin			
			Asin					
				cos	Cos			
			Acos					
				atan	Atan			
		tan	Tan					
		reciprocal	Reciprocal					
					Square			
			Sqrt					
					Rsqrt			
			Exp					
					Log			
BIAS	√							
NOOP								
THRESHOLD	√							
HARDSIGMOID								
EMBEDDING	√	√	√					
INSTANCENORM				√				
MVN	√							

TENGINE	Caffe	MXNet	TensorFlow	TF-lite	ONNX	Darknet
ABSVAL	√					
CAST	√					
HARDSWISH					√	
INTERP	√	UpSampling	Upsample			
SELU						
ELU	√	LeakyReLU	ELU	√		
BROADMUL			broadcast_mul			
LOGICAL			LOGICALOR			
				LOGICALAND		
GATHER			GATHER	√		
TRANSPOSE				√	TRANSPOSE	√
COMPARISON		Equal	EQUAL			
				Greater	GREATER	
		GreaterEqual	GREATER_EQUAL			
				Less	LESS	
		LessEqual				
					LESS_GREATER	
SPACETODEPTH			SPACE_TO_DEPTH			
DEPTHTOSPACE					DEPTH_TO_SPACE	
REVERSE		ReverseV2	REVERSE_V2			
SPARSETODENSE				√	SPARSE_TO_DENSE	
CEIL		√	CEIL			
SQUAREDDIFFERENCE				√	SQUARED_DIFFERENCE	
ROUND		√	ROUND			
ZEROSLIKE						
CLIP	Clip		Clip			
POWER	Power					
TILE	Tile					
L2NORMALIZATION					L2_NORMALIZATION	
L2POOL			L2_POOL_2D			
RELU1					RELU_N1_TO_1	
LOGSOFTMAX			LOG_SOFTMAX			
FLOOR				Floor		
REDUCEL2			√			
UNSQUEEZE						√