Tengine
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 是 OPEN AI LAB 开发的嵌入式高性能轻量级深度学习推理框架，目标是提供 Arm 嵌入式平台最佳的深度学习模型部署体验产品构成与主要功能，图 1 是框架架构图。

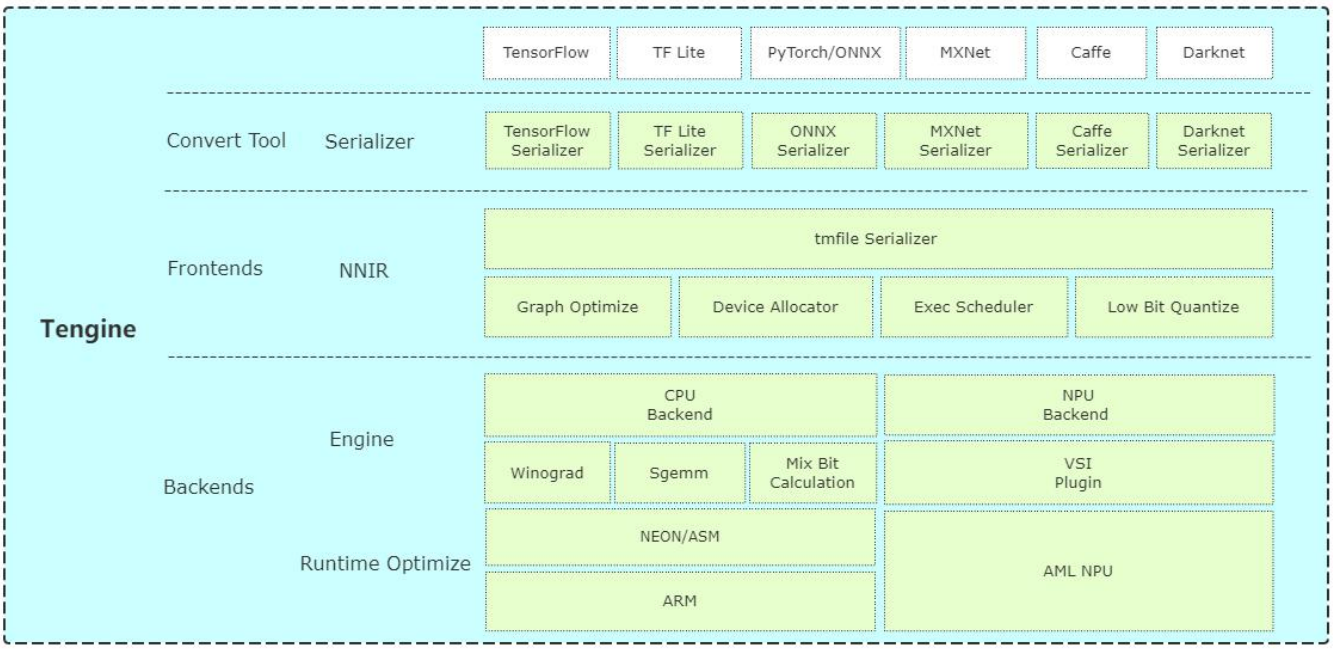

图 1 Tengine Architecture

1.1.1 Tengine

Tengine 是开源框架，模块化设计，在计算时只依赖于 C 标准库。

1.2 产品特点

1) 针对不同 CPU 微架构以及 SoC 系统高度优化 CPU 性能，针对 Khadas A331D 平台，适配 Arm Cortex-A73/A53；
2) 运行 Caffe/MXNet/TensorFlow/TFLite/ONNX/DarkNet 模型文件，需先转换为 tmfile 模型；
3) 针对内存优化设计的框架和算子接口定义，极大减少了内存占用；
4) 核心模块不依赖于第三方库，只依赖于系统 C 标准库。

2 支持范围

2.1 硬件支持

2.1.1 CPU 的支持

支持 Armv8a

2.1.2 NPU 的支持

- 支持 Amlogic NPU

2.2 操作系统支持

- Android 9
- Ubuntu 20.04

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 模型支持

2.4.1 Caffe 模型支持

AlexNet	Faster_rcnn	GoogleNet	Inception_v3	Inception_v4
LighteCNN	Mobileface	MobileNet_v1	MobileNet_SSD	MTCNN

ResNet50	SqueezeNet	SSD	VGG16	VGG19
YOLOv2	yufacedetect	MobileNet_v2	MobileNet_v3	ShuffleNet_1xg3
MnasNet	ShuffleNet_v2			

2.4.2 ONNX 模型支持

SqueezeNet	MobileNetV3	ShuffleNetV2		

2.4.3 MXNet 模型支持

MobileFaceNet	MobileNet	SqueezeNet	MobileNet_v2	Inception_v3
ResNet50	VGG16	AlexNet	ResNet18_v2	

2.4.4 TensorFlow 模型支持

Inception_v3	Inception_v4	MobileNet_v1	MobileNet_v2	ResNet50
ResNet_v1	ResNet_v2	SqueezeNet	DenseNet	NasNet
MobileNet_v1_0.75	Inception_ResNet_v2			

2.4.5 TFLite 模型支持

Squeezenet	Inception_v4			

2.4.6 DarkNet 模型支持

YOLOv2	YOLOv2 tiny	YOLOv3	YOLOv3 tiny	

2.4.7 模型解密支持

支持对加密后的 tmfile 模型文件的解密操作。

具体产品规格，请参考《Tengine Encrypt Tool Spec》

2.5 计算模式支持

Float32、Hybrid Int8。

2.6 调度策略支持

2.6.1 多线程支持

支持指定 CPU 多线程运算，支持 CPU 亲和性绑定。

2.6.2 异构计算支持

支持 Arm CPU/NPU 异构计算。

2.7 工具支持

2.7.1 Convert Tool

模型转换工具支持转换 Caffe/MXNet/TensorFlow/TFLite/ONNX/DarkNet 类型的模型到 tmfile 格式的模型。具体产品规格，请参考《Tengine Convert Tool Technical Spec》。

2.7.2 Netron

Netron 实现可视化 tmfile 网络模型的功能。

3 其他软件产品的依赖

Tengine 不依赖任何第三方库。但不排除 Tengine 的部分工具、Tengine 的用户使用场景需要依赖部分第三方库。

4 性能数据

（待 QA 补充）

附录 1 Tengine 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				
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

Tengine	Caffe	MXNet	TensorFlow	TF-Lite	ONNX	DarkNet
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			
			Logsum			
			Logsumexp			
Argmax			√			
Argmin			√			
Topkv2			√			
Maximum			√		Max	
Minimum			√			
Addn		add_n				
Swapaxis		√				
Gru		RNN	√			
Upsample	√	UpSampling				upsample
Shufflechannel	√					
Resize	√		ResizeNearestNeighbor			
			ResizeBilinear			
Spacetobatchnd			√			
Batchtospacend			√			

Tengine	Caffe	MXNet	TensorFlow	TF-Lite	ONNX	DarkNet
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	√					
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		

Tengine	Caffe	MXNet	TensorFlow	TF-Lite	ONNX	DarkNet
Ceil			√	CEIL		
Squareddifference			√	SQUARED_DIFFERENCE		
Round			√	ROUND		
Zeroslike						
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					√	
Relu10				√		
Mish						√
Scatter					√	
Shape					√	
Where					√	

注：每列表示此框架支持算子名称；“√”表示此框架下相同功能算子与 Tengine 名称相同。