Tengine 技术规格书

文档版本 0.4

发布日期 2019-08-27

OPEN AI LAB

OPEN AI LAB

变更记录

日期	版本	说明	作者
2019-06-11	0.1	初版	rui
2019-08-15	0.2	更新算子列表	rui
2019-08-27	0.3	更新模型和算子列表	CMeng
2019-11-08	0.4	更新算子列表	CMeng

OPEN AI LAB

目录

OPEN AI LAB

OPEN AI LAB

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。

1.2 产品构成与主要功能

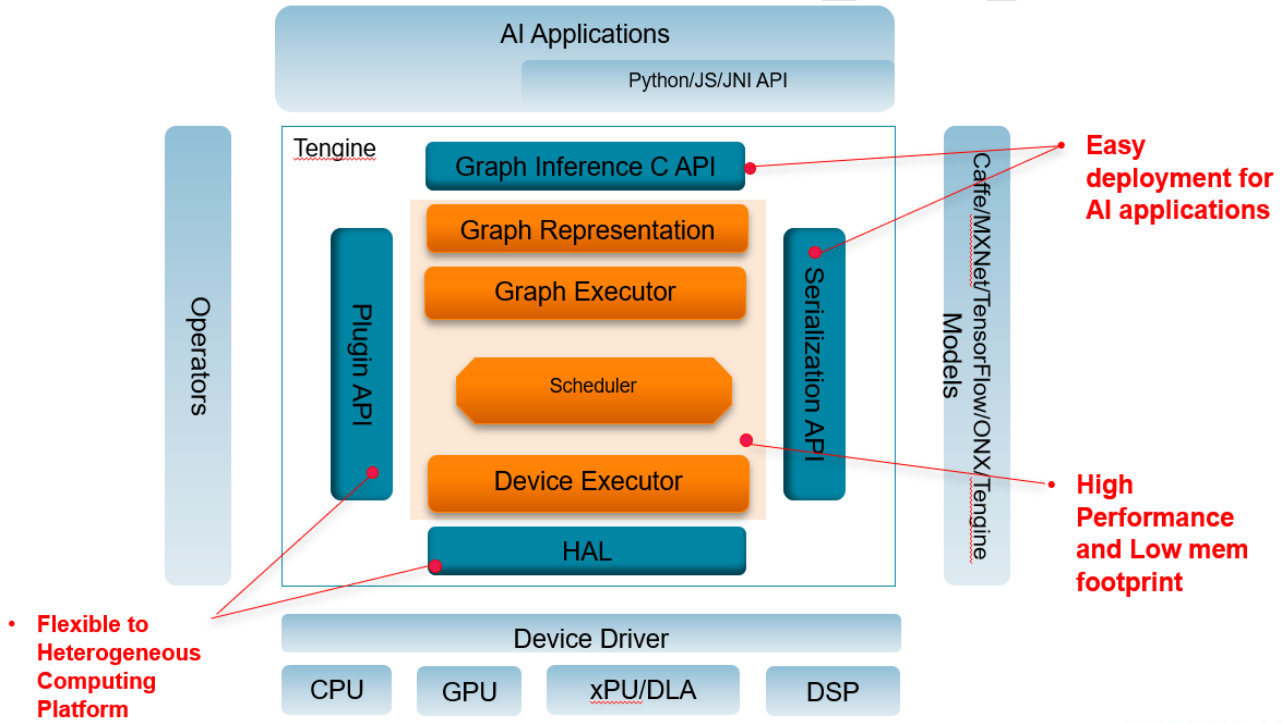

图 1

1.2.1 Tengine

Tengine 是开源框架，模块化设计，在计算时只依赖于 C/C++库[1]。

[1] 注：Tengine 在 x86 上运行时， 为了更好的性能， 需要 OpenBlas 库支持；以及在解析其他模型文 件时，需要相应的解析库支持，相关信息请参见 3.0 章节。

1.2.2 HCL

HCL 是针对 Arm CPU 开发的高性能 NN 推理计算库。

1.3 产品特点

1) 针对 Arm CPU 微架构以及 SoC 系统高度优化的 CPU 性能，适配 Arm Cortex-A7/A17/A53/A72/A73/A55/A76；
2) 可以直接加载 Caffe/Mxnet/Tensorflow/TF-Lite/ONNX/DarkNet 模型文件，而不需要事先转换；
3) 基于 Caffe/Tensorflow/Mxnet 开发的应用，仅需重新编译就可以利用 Tengine 的加速性能；
4) 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；
5) 模块化设计：通过扩展接口可以定义和实现新的 operator；
6) 核心模块不依赖于第三方库，只依赖于系统 C/C++运行库；
7) 支持多设备异构计算：可以利用 CPUs，GPUs 和 DLAs 进行运算；
8) 支持层融合、8bit 量化等优化策略。

2 支持范围

2.1 硬件支持

2.1.1 CPU 的支持

支持所有 Armv7a，Armv8a CPU

Armv7a: Cortex-A5,A7,A15,A17,A32
Armv8a: Cortex-A35,A53,A57,A72,A73
Armv8.2a: Cortex-A55,A75,A76

2.1.2 GPU 的支持

支持 Arm Mali-T6xx, T8xx, G 系列 GPU，需支持 OpenCL 驱动

OPEN AI LAB

2.1.3 DLA 的支持

- 支持 Arm 中国周易 AIPU
- 支持海思 NNIE，Hi3559a，Hi3519a, Hi3516cv500, Hi3516dv300
- RK3399Pro NPU

2.2 操作系统支持

- Android 5.1 以上所有版本
- Linux：Fedora, Ubuntu, Debian

2.3 算子支持

2.3.1 Tengine 算子支持

- Tengine 支持算子总计 74 个：

Absval, Accuracy, Addn, ArgMax, ArgMin, BatchNormalization, BatchToSpaceND, Bias, Cast, Concat, Const, Convolution, Crop, Deconvolution, DetectionOutput, DetectionPostProcess, Dropout, Eltwise, Elu, Embedding, Expanddims, Flatten, FullyConnected, Fused.BNScaleReLu, GRU, Gemm, Generic, HardSwish, Hardsigmoid, InputOp, InstanceNorm, Interp, LRN, LSTM, Logistic, MVN, Maximum, Minimum, Noop, Normalize, PReLU, Pad, Permute, Pooling, PriorBox, Psroipooling, RNN, ROIPooling, RPN, ReLu, ReLu6, Reduction, Region, Reorg, Reshape, Resize, Resize, Roialign, Scale, Selu, ShuffleChannel, Sigmoid, Slice, Softmax, SpaceToBatchND, Split, Squeeze, StridedSlice, SwapAxis, Tanh, Threshold, TopKV2, Unary, Upsample。

- Tengine 支持 Caffe 算子总计 49 个：

AbsVal, Accuracy, BatchNorm, Bias, Clip, Concat, Convolution, ConvolutionDepthwise, Crop, Data, Deconvolution, DepthwiseConvolution, DetectionOutput, Dropout, ELU, Eltwise, Embedding, Flatten, InnerProduct, Input, Interp, LRN, MVN, Normalize, PReLU, Permute, Pooling, Power, PriorBox, ROIPooling, RPN, ReLU, ReLU6, Reduction, Region, Reorg, Reshape, Resize, Scale, ShuffleChannel, Sigmoid, Slice, Softmax, SoftmaxWithLoss, Split, TanH, Threshold, Tile, Upsample。

- Tengine 支持 TensorFlow 算子总计 66 个：

OPEN AI LAB

Acos, Add, AddN, ArgMax, ArgMin, Asin, Atan, AudioSpectrogram, AvgPool, BatchToSpaceND , Cast , ComposedBN, ConcatV2, Conv2D, Conv2DBackpropInput, Cos, DecodeWav, DepthwiseConv2dNative, Dropout, Embedding , Exp, ExpandDims, FIFOQueueV2, Flatten, Floor, FusedBatchNorm, GRU, LRN, LSTM, Log, MatMul, MaxPool, Maximum, Mean, Mfcc, Minimum, Minimum, MirrorPad, Mul, Pad, Pow, RNN, RealDiv, Reciprocal, Relu, Relu6, Reshape, ResizeBilinear, ResizeNearestNeighbor, ReverseV2, Rqrt, Rsqrt, Rsqrt, Sigmoid, Sin, Softmax, SpaceToBatchND, Split, Sqrt, StridedSlice, Sub, Sum, Tan, Tanh, TopKV2, Unary 。

- Tengine 支持 MXNet 算子总计 41 个:

Activation, BatchNorm, Concat, Convolution, Copy, Copy, Crop, Crop, Deconvolution, Dropout, Embedding, Flatten, FullyConnected, InstanceNorm, LeakyReLU, Pooling, RNN, Reduction, Reshape, SoftmaxActivation, SoftmaxOutput, SwapAxis, UpSampling, UpSampling, _contrib_PSROIPooling, _contrib_ROIAlign, _minus_scalar, _mul_scalar, abs, add_n, atan, ceil, clip, cos, elemwise_add, floor, neg, reciprocal, sin, tan, transpose。

- Tengine 支持 ONNX 算子总计 23 个:

Add, AveragePool, BatchNormalization, Clip, Concat, Conv, Div, Dropout, Elu, Flatten, Floor, Gemm, GlobalAveragePool, HardSwish, LeakyRelu, MaxPool, Mul, PRelu, Relu, Reshape, Softmax, Transpose, Upsample。

- Tengine 支持 Darknet 算子总计 8 个:

convolutional, maxpool, region, reorg, route, shortcut, upsample, yolo。

- Tengine 支持 TF-Lite 算子总计 18 个:

ADD, AVERAGE_POOL_2D, CONCATENATION, CONV_2D, DEPTHWISE_CONV_2D, ELU, L2_NORMALIZATION, L2_POOL_2D, LOGISTIC, LOG_SOFTMAX, MAX_POOL_2D, RELU_N1_TO_1, RESHAPE, RESIZE_NEAREST_NEIGHBOR, SOFTMAX, SQUEEZE, STRIDED_SLICE, TFLite_Detection_PostProcess。

卷积计算方法包括:

- Direct Convolution
- Winograd Convolution
- Gemm Convolution

OPEN AI LAB

2.4 模型支持

2.4.1 Caffe 模型支持

Alexnet	faster_rcnn	googlenet	inception_v3	inception_v4
lighten_cnn	mobileface	Mobilenet_v1	mobilenet_ssd	mtcnn
resnet50	squeeznet	ssd	vgg16	vgg19
yolov2	yufacedetect	Mobilenet_v2	Mobilenet_v3	Shufflenet_1xg3
Mnasnet	Shufflenet_v2	Lightcnn		

2.4.2 ONNX 模型支持

squeeznet				

2.4.3 Mxnet 模型支持

mobileface	mobilenet	squeeznet	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	ResnNet50
ResnNet_v1	ResnNet_v2	squeeznet	densenet	nasnet
Mobilenet_v1_0.75	Inception_resnet_v3			

OPEN AI LAB

2.4.5 Tensorflow Lite 模型支持

ResnNet_v2	inception_v3	squeeznet	Mobilenet_v1	Mobilenet_v2
Inception_v3	Inception_v4	mobilenet_ssd	mobilenet_quant	detect

2.4.6 Darknet 模型支持

Yolov2	Yolov2 tiny	Yolov3	Yolov3 tiny	

2.4.7 模型加密支持

支持二进制加密

2.5 计算模式支持

Float32, Float16, int8, uint8

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

支持在 CPU，GPU，AIPU 等多个 devices 上进行多线程计算

2.6.2 异构计算支持

支持 Arm CPU/GPU 异构计算

OPEN AI LAB

2.7 工具支持

2.7.1 Tengine 模型转换工具

支持将 Caffe/Onnx/Mexnet/Tensorflow/Tensorflow Lite/Darknet 的模型文件转换成 Tengine 的模型文件。

2.7.2 Tengine 模型量化工具

- 支持将 Tengine 的 FP32 模型量化成 FP16 的模型

2.7.3NCHW Post trainging tool

Post Training Tool 主要功能将 Tengine FP32 model 转化为 Tengine End2End INT8 model.

2.8 其他框架 API 的支持

2.8.1AndroidNN API 支持

AndroidNN Operator	Tensor Type Supported
ANEURALNETWORKS_CONV_2D	FLOAT32,QUANT8_ASYMM
ANEURALNETWORKS_DEPTHWISE_CONV_2D	FLOAT32,QUANT8_ASYMM
ANEURALNETWORKS_AVERAGE_POOL_2D	FLOAT32,QUANT8_ASYMM
ANEURALNETWORKS_CONCATENATION	FLOAT32,QUANT8_ASYMM
ANEURALNETWORKS_SOFTMAX	FLOAT32,QUANT8_ASYMM
ANEURALNETWORKS_RESHAPE	FLOAT32,QUANT8_ASYMM
ANEURALNETWORKS_SQUEEZE	FLOAT32,QUANT8_ASYMM

OPEN AI LAB

3 其他软件产品的依赖

3.1 Protobuf

Tengine 需要 3.0.0 version 以上的版本支持，用于 Tensorflow/Caffe 模型文件解析使用。

3.2 Openblas

当在 X86 平台上运行 Tengine, 如需加速，需要 OpenBlas 库支持。

3.3 ACL

Tengine 是通过调用 Arm Compute Library（ACL）进行 GPU 加速，使用的 ACL 版本为 19.02。

4 性能数据

Tengine 支持各类 Arm SoC 平台，通过对芯片微构架针对性优化，充分挖掘出芯片的潜力，将性能和硬件利用率提到最高。

RK3399 平台包括 2 核 Cortex-A72 1.8GHz，4 核 Cortex-A53 1.4GHz 和 Mali-T860MP4 GPU

在 RK3399 上测试性能数据如下（单位：ms）：

	FP32				INT8			
	1 x A72	2 x A72	1 x A53	4 x A53	1 x A72	2 x A72	1 x A53	4 x A53
MobileNet v1	**111.5**	**65.7**	**224.6**	**75.9**	**75.2**	**45.3**	**162.7**	**60.2**
SqueezeNet	**60.8**	**42.7**	**123.5**	**57.6**	**55.2**	**36.9**	**126.1**	**54.8**

RK3288 平台包括 4 核 Cortex-A17 1.8GHz 和 Mali-T760MP4 600MHz GPU

Tengine 在 RK3288 上测试性能数据如下（单位：ms）：

	FP32		INT8	
	1 x A17	4 x A17	1 x A17	4 x A17

OPEN AI LAB

MobileNet v1	**196.1**	**67.2**	**103.6**	**37.6**
SqueezeNet	**110.8**	**46.1**	**80.4**	**32.2**

Allwinner R40 平台包括 4 核 Cortex-A7 1.2GHz

Tengine 在 Allwinner R40 上测试性能数据如下（单位：ms）：

	FP32		INT8	
	1 x A7	4 x A7	1 x A7	4 x A7
MobileNet v1	**710.5**	**210.1**	**412.6**	**138.8**
SqueezeNet	**380.2**	**143.7**	**316.0**	**117.2**

注：以上数据基于 Tengine V1.10.0 的 测试结果。

OPEN AI LAB