

新版工具链yaml配置文件参数详解

1 引言

目前地平线浮点转换工具链需要依据 yaml 配置文件实现模型转换，而某些参数如果配置有误将会导致最终模型的性能/精度不理想，甚至直接转换失败，本文将针对 yaml 文件中的相关配置项进行说明，作为工具链文档的额外补充。

 **Note:** 随着工具链版本迭代，yaml 参数可能会有细微调整，本文以 **v1.9.4** 版本 MobileNet-v1 为例进行说明

2 概述

整个 yaml 文件包含 **4 个必选参数组** (`model_parameters`、`input_parameters`、`calibration_parameters`、`compiler_parameters`) 和 **1 个可选参数组** (`custom_op`)，每个参数组下也区分必选和可选参数（可选参数默认隐藏），具体要求和填写方式请参考工具链文档。其中以下为几个填写注意点：

1. 当模型为多输入模型时，请将可选参数 `input_name`，`input_shape` 显式写出，以免造成参数对应顺序上的错误；
2. 当 `input_type_rt` 配置为 `nv12` 时，模型H/W维度不支持奇数输入大小，例如不支持 1x255x255x3。

3 参数详解

本小节针对各部分参数进行分类说明，主要包括 **性能影响参数**、**精度影响参数**、**自定义算子参数** 三类。关于参数的具体描述或取值范围，可参考对应的工具链文档。

3.1 性能影响参数

关于配置文件中与性能相关参数总结（包含未在示例中暴露的高阶参数）如下表所示：

	A	B	C
1	参数组	参数名称	性能影响
2	model_parameters	layer_out_dump	如果配置为 True，则会为每一个卷积算子增加输出节点，主要用于工具进行精度分析，但会显著降低性能； 正常转换请配置为 False
3	input_parameters	input_type_layout	X3 的 BPU 硬件只支持 NHWC 的运行模型，所以对 input_type_t rgb/bgr 的模型来说，如果 input_type_layout 配置为 NCHW，则 入一个 transpose 节点实现数据转换，此时会有额外的性能耗时。 建议在数据通路中直接传入 NHWC 的数据，这样 BPU 能够直接进行
4	compiler_parameters	compile_mode	通常情况下配置为 latency 以获取高效性能优化，而当带宽遇到瓶 bandwidth，此时性能会有所下降
5		debug	配置为 True 将打开编译器的 debug 模式，可以在静态性能评估文 子的耗时信息，如图1所示，一般用于性能分析，但因为模型本身 数信息，所以也会对最终性能造成少量影响； 在实际产品交付阶段，请配置为 False
6		core_num	配置为2时可同时调用两个核运行，降低单帧推理延迟，但同时也 吐率； 默认该参数值为1即可
7		optimize_level	用于生产或验证性能时，务必配置成O3等级优化才能保证最佳性能
8	calibration_parameters	可选参数： max_time_per_fc	该参数与优先级抢占相关，默认值为0（即最高优先级） 一般不需要配置添加此参数
9		可选参数： run_on_bpu	如果您对性能有较高的要求，愿意以一些量化损失为代价，则可以 运行在BPU上； 一般不需要配置添加此参数
10		可选参数： run_on_cpu	CPU上提供的是float精度计算，如果您确定某些算子需要在CPU上 运行在CPU上，但性能会有所下降； 一般不需要配置添加此参数

Summary

Temporal Statistics

Layer Details

Layer Details:

layer	original ops	computing time (us)
HZ_PREPROCESS_FOR_data	903168	25
Conv_0-conv	21676032	56
Conv_3-conv	7225344	14
Conv_6	12845056	25
Conv_8-conv	38535168	75
Conv_11-conv	5419008	12
Conv_14	14450688	28
Conv_16-conv	21676032	42
Conv_19-conv	8128512	18
Conv_22	21676032	42
Conv_25-conv	21676032	42

图1 Layer Details信息

3.2 精度影响参数

关于配置文件中与精度相关参数总结（包含未暴露的参数）如下表所示：

	A	B	C
1	参数组	参数名称	精度影响
2	input_parameters	input_type_train	该参数为模型训练时的数据类型，如果该参数配置错误，会造成一定（例如：误将rgb配置为bgr）
3		norm_type/ mean_value/ scale_value	设置模型输入归一化操作； 如果在yaml文件中配置了归一化操作，则在处理校准数据和验证/测试需再做归一化，重复操作会引入精度损失
4		cal_data_dir	样本选取不当，对精度有较大的影响。关于标定样本选取，需注意以 1. 图片数量一般控制在20~100张即可 2. 不选择质量差的图片，包括纯黑纯白、过曝光、模糊等图片 3. 选择的图片范围尽可能覆盖典型场景
5		calibration_type	不同校准算法在不同模型上的精度表现可能会有差异； 建议大家优先尝试 default 参数，另外值得注意的是：如果是加载通感知训练）模型，此参数需要配置为：load
6		可选参数： max_percentile	建议大家尝试不同参数值情况下进行精度调优； 此参数需要搭配 max 校准使用，常用的参数值有： 0.99999/0.99995/0.99990/0.99950/0.99900
7	calibration_parameters	可选参数： per_channel	可与任意一种校准方式组合使用，如 kl、max、max+max_percent

3.3 自定义算子参数

此部分参数针对您有自定义算子行为的需求进行配置，目前我们已开放部分CPU算子自定义，更多关于如何自定义算子请参考文章[自定义算子](#)这一小节，相关参数配置说明如下：

	A	B	C
1	参数名称	描述（功能）说明	取值范围
2	custom_op_method	自定义算子策略选择	register
3	op_register_files	自定义算子的Python实现文件名称	名称自拟，如果多个文件用；进行分隔
4	custom_op_dir	自定义算子的Python实现文件存放路径	名称自拟，设置路径，请使用相对路径