IPU 介绍

文档版本号：v1.0
发布日期：2020-1-10

商标声明

、全志和其他全志商标均为珠海全志科技股份有限公司的商标。本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受全志公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，全志公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保

前言

概述

本文档为基于 NNTurbo 开发算法模型的工程师提供开发指导。

产品版本

与本文档对应的产品版本。

产品名称	产品版本
V833	NNTurbo V1.0

读者对象

本文档（本指南）主要适用于以下工程师：

使用 NNTurbo 进行开发的软件/算法开发工程师

名词解释

API——应用程序接口,全称：Application Programming Interface

SDK——软件开发包，全称：Software Development Kit

NNTurbo——AW 中间件总称

修订记录

版本号	修订日期	修订内容
IPU 介绍 V1.0	2020-1-10	第一次发布。

1. IPU 介绍

IPU 中全志智能处理器的简称，是全志 SoC 针对深度学习卷积神经网络进行加速处理的硬件单元，支持目前流行的卷积神经网络网络及算子。软件 API 开发是基于全志的 NNTurbo 软件栈进行，可参考《NNTurbo API 指南》。

2. IPU 规格

2.1. 数据格式

NNTurbo 目前只支持 INT8（有符号 8 位）运算，主要是针对权重，特征图以及输入图像。

2.2. 算子支持

算子	描述
conv(卷积)	卷积算子，bias 建议按通道偏置的模式，按 layer 和 point 也可以。
pooling(池化)	池化算子，池化核 <= 7x7。
eltwise(元素对位运算)	支持加法，乘法，乘加运算。
lrn(局部响应归一化)	暂不使用。
inner_product(内积运算)	全连接算子。
activation(激活运算)	支持 relu，prelu。
bn(批归一化)	采用通道模式进行批归一化(Batch Normalization)操作。
conv_act_pool (卷积、激活、池化复合运算)	卷积，激活，池化的融合算子，数据会在模块内部 bypass，实现减少带宽的目的，限制与单个算子的限制相同。
conv_act (卷积、激活复合运算)	卷积，激活融合算子，数据会在模块内部 bypass，实现减少带宽的目的，限制与单个算子的限制相同。
conv_act_eltwise (卷积、激活、元素对位复合运算)	卷积，激活，对位运算算子，数据会在模块内部 bypass，实现减少带宽的目的，限制与单个算子的限制相同。（此算子主要用于残差网络运算。）
conv_bn_pool （卷积、bn、池化复合运算）	卷积，batch normalization，池化融合算子，数据会在模块内部 bypass，实现减少带宽的目的，限制与单个算子的限制相同。
conv_act_bn （卷积、激活、bn 复合运算）	注意 prelu 与 bn 不能同时使用。
conv_act_bn_pool	注意 prelu 与 bn 不能同时使用。

（卷积、激活、bn、池化复合运算）	

2.3. 框架支持

当前只支持 TensorFlow 的模型，其它框架的模型需要转换成 TensorFlow 模型。

2.4. 注意事项

支持数据及权重为 8bit(int8)。卷积，池化算子的 pad 操作，只支持 pad 1 个 byte，默认值是 0,可分别单独设置上下左右。

受 IPU 单元 Buffer 大小约束，参与运算的特征图与权重占内存的总大小不能超过 256KB,若超出，则需保证特征图+8 通道（8 kernels）权重值大小超过 256KB，工具会对权重按 8 通道（8 kernels）切分，以上按 int8 计算。

卷积核的通道数为 32 的倍数时，效率最高；其它也能运行。

3. IPU 工具链

IPU 提供的软件工具链总称是 NNTurbo，其包含两个部分：

- 在线部分：嵌入式运行库，目前支持 linux 及全志 Tina 系统。
- 离线部分：将模型文件转换为 IPU 可计算的 8bit 定点数据，同时进行量化。

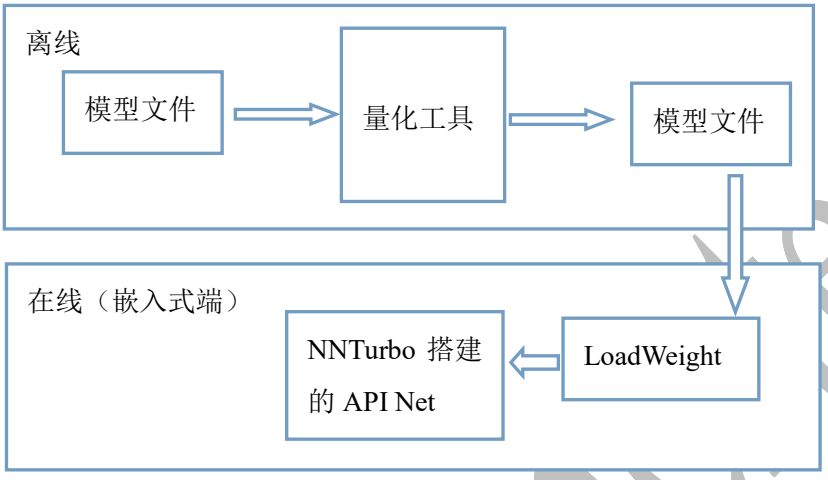
注意，目前工具只支持 tensorflow 的模型转换。

4. 开发流程

4.1. TensorFlow 模型

关于 TensorFlow 模型，目前只能支持 2.2 列表中支持的算子，对于不支持的算子，用户需自行对网络进行切割然后转换。

TensorFlow 模型转换流程如下：



4.2. 模型转换注意事项

该工具只支持 TensorFlow 模型的转换，模型存储建议 tensorflow-1.13.1 版本。

模型转换及量化工具(/tools/main.exe)调用之前需要配置参数列表(/tools/config.ini)，配置完成后运行该工具，然后点击 optimizer,运行无误后，再点击 quatizer。量化成功之后会将量化好的模型文件存放在(/tools/bin/)路径下。模型文件名格式为 data-(模型文件生成时间).bin。

Config.ini 文件配置参数说明如下。

model_path	tensorflow 模型存放路径（只支持 ckpt 格式）。
in_names	tf 模型输入 tensor_names 的列表。
out_names	tf 模型输出 tensor_names 的列表。
images_path	量化所需图片的存放路径（不能包含中文字符）。
num	量化图像的数目，必须与 tf 模型对应训练时的 batch_size 相同。
filter	量化图片的图像格式。
image_scale	量化图像的缩放尺度，只支持整数倍放大。

前向推断时需要用户将输入数据预处理后自行量化为 int8 数据，输入图像的排列格式为(HWC)。