AW NPU 常见网络性能分析数据

版本号: 1.0

发布日期: 2021.07.21

版本历史

版本号	日期	制/修订人	内容描述
1.0	2021.07.21	PDC	NPU 工具安装说明。

目　录

1 前言

1.1 读者对象

本文档（本指南）主要适用于以下人员：

- 技术支持工程师
- 软件开发工程师
- AI 应用案客户

1.2 约定

1.2.1 符号约定

本文中可能出现的符号如下：

警告

警告

技巧

1. 技巧
2. 小常识

说明

说明

2 正文

2.1 NPU 开发简介

- 支持 int8/uint8/int16 量化精度，运算性能可达 1TOPS.
- 相较于 GPU 作为 AI 运算单元的大型芯片方案，功耗不到 GPU 所需要的 1%.
- 可直接导入 Caffe, TensorFlow, Onnx, TFLite，Keras, Darknet, pyTorch 等模型格式.
- 提供 AI 开发工具：支持模型快速转换、支持开发板端侧转换 API、支持 TensorFlow, TF Lite, Caffe, ONNX, Darknet, pyTorch 等模型.
- 提供 AI 应用开发接口：提供 NPU 跨平台 API.

2.2 开发流程

NPU 开发完整的流程如下图所示：

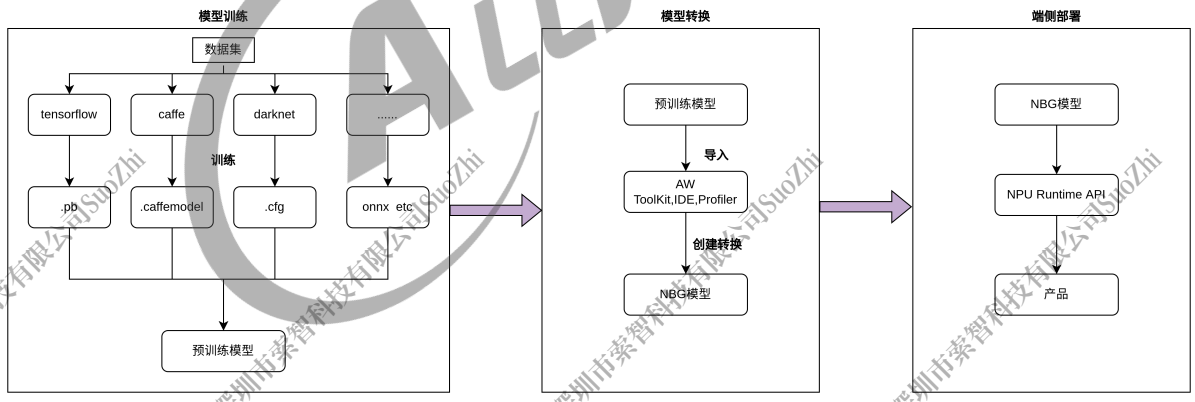

图 2-1：npu_1.png

2.3 常见网络 benchmark

NPU Performance			
Case	FPS @ 500MHz	BW(MB/s)@ 500MHz	BW(MB/Inf)
inception_v1(224x224)	48.25	1213.82	25.16
inception_v3_224(224x224)	31.24	1425.55	45.63
inception_v3_299(299x299)_tf	18.72	1392.02	74.45
mobilenet_v1(224x224)	141.00	2134.70	15.14
mobilenet_v2(224x224)_tflite	124.54	2260.47	18.15
mobilenet_v2(224x224)_onnx	139.71	2451.92	17.53
resnet_v1_50(224x224)	27.71	1440.77	51.99
retinanet_resnet_50(320x320)	6.48	1746.42	269.27
vgg16(224x224)	8.87	1835.00	206.89
yolo_v2_voc(416x416)	15.89	1315.91	82.79
yolov3(416x416)	3.78	1311.20	346.60
yolov4(416x416)	4.39	1598.47	363.72
yolov4-tiny(416x416)	50.08	1526.75	30.48
yolov5(416x416)	10.73	1420.09	132.25
yolov5s(640x640)	5.51	1379.77	250.60
lenet(28x28)	2356.66	871.96	0.37
AW_person_detection_darknet	55.40	896.92	16.19
AW_Face_detection_onnx	61.11	2028.12	33.19
inception_v4_299(299x299)	9.26	1684.87	182.03
resnet18(224x224)	72.57	1008.72	13.90
googlenet(224x224)	50.63	1095.86	21.58
alxnet(224x224)	36.29	1764.61	48.63
inception_v3_299(299x299)_onnx	18.67	1388.02	74.35

图 2-2: NPU benchmark

以上数据是裸机程序跑网络的数据，并未考虑到方案中的其它应用。

2.4 内存分析数据

方案应用场景中的内存消耗数据分析.

代码和数据部分的占用，包括 KMD 和 UMD 本身占用的空间大小, 大约 180k.

	text	data	bss	总计
内核态	55164	920	388	**56472**
用户态	99739+22656	604+484	388+72	**123943**
总计	99739+22656+55164=**177559**	604+484+920=**2008**	388+72+388=**848**	**180415**

图 2-3: code 占用大小

Yolov3 模型的内存数据统计，运行时消耗约 48M 内存。

	total video memory	total system memory	viplite driver code size	total
大小	48460032	81500	180415	48721947
占比	99.46%	0.17%	0.37%	100%

图 2-4: yolov3 内存统计

yolov3-tiny 模型的内存数据统计，运行时消耗月 6.8M 内存。

	total video memory	total system memory	viplite driver code size	total
大小	6710784	20596	180415	6911795
占比	97.092%	0.307%	2.61%	100%

图 2-5: yolov3-tiny 内存统计

帧率，带宽等数据待补充.

3 结束

著作权声明

本文档及内容受著作权法保护，其著作权由珠海全志科技股份有限公司（“全志”）拥有并保留一切权利。

本文档是全志的原创作品和版权财产，未经全志书面许可，任何单位和个人不得擅自摘抄、复制、修改、发表或传播本文档内容的部分或全部，且不得以任何形式传播。

商标声明


、 

、 
（不完全列举）均为珠海全志科技股份有限公司的商标或者注册商标。在本文档描述的产品中出现的其它商标，产品名称，和服务名称，均由其各自所有人拥有。

免责声明

您购买的产品、服务或特性应受您与珠海全志科技股份有限公司（“全志”）之间签署的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所购买或使用的范围内。使用前请认真阅读合同条款和相关说明，并严格遵循本文档的使用说明。您将自行承担任何不当使用行为（包括但不限于如超压，超频，超温使用）造成的不利后果，全志概不负责。

本文档作为使用指导仅供参考。由于产品版本升级或其他原因，本文档内容有可能修改，如有变更，恕不另行通知。全志尽全力在本文档中提供准确的信息，但并不确保内容完全没有错误，因使用本文档而发生损害（包括但不限于间接的、偶然的、特殊的损失）或发生侵犯第三方权利事件，全志概不负责。本文档中的所有陈述、信息和建议并不构成任何明示或暗示的保证或承诺。

本文档未以明示或暗示或其他方式授予全志的任何专利或知识产权。在您实施方案或使用产品的过程中，可能需要获得第三方的权利许可。请您自行向第三方权利人获取相关的许可。全志不承担也不代为支付任何关于获取第三方许可的许可费或版税（专利税）。全志不对您所使用的第三方许可技术做出任何保证、赔偿或承担其他义务。