A733 NPU 常见网络性能
测试报告

版本号: 1.1
发布日期: 2025.05.10

深圳市宇芯数码技术有限公司yxq_01

版本历史

版本号	日期	制/修订人	内容描述
1.0	2024.11.13	AWA1382	初始版本
1.1	2025.05.10	AWA1382	补充支持 Android13、Linux buildroot 系统

深圳市宇芯数码技术有限公司yqq_01

版权所有 © 珠海全志科技股份有限公司。保留一切权利

目　录

1 前言 **1**
1.1 文档简介 1
1.2 目标读者 1
1.3 适用范围 1
1.4 文档约定 1
1.4.1 术语，缩略语及概念 1
1.4.2 标志说明 2

2 环境搭建 **3**

3 测试步骤 **5**

4 公开模型的性能测试 **7**

5 公开模型的资源情况 **8**

6 附录 **9**

深圳市宇芯数码技术有限公司yfzy_01

版权所有 © 珠海全志科技股份有限公司。保留一切权利

1 前言

1.1 文档简介

介绍 NPU 模块的常见网络性能的测试报告，为 NPU 驱动、算法开发提供参考。

1.2 目标读者

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

1.3 适用范围

硬件平台：A733 平台

软件平台：Android13、Android 15 系统、Linux buildroot 系统

1.4 文档约定

1.4.1 术语，缩略语及概念

术语	解释说明
NPU	Neural Network Processing Unit，神经网络处理器。
VIPLite	VIPLite 版本 NPU 驱动。
Unified	Unified 版本 NPU 驱动。
带宽	发送信号中含有的有效成分的效率范围，可以用来标识信号传输的数据传输能力，简写 BW。

版权所有 © 珠海全志科技股份有限公司。保留一切权利

1.4.2 标志说明

本文中出现的符号如下：

说明

为准确理解文中指令、正确实施操作而提供的补充或强调信息。

深圳市宇芯数码技术有限公司yxy_01

版权所有 © 珠海全志科技股份有限公司。保留一切权利

2 环境搭建

一、编译 NPU 驱动并传到板端加载（可选，公版方案中已默认编译，则无需操作）：

```
#加载VIPLite版本NPU驱动，当前建议使用此驱动。
insmod vipcore.ko
#加载Unified版本NPU驱动，两个版本二选一，不能同时加载。
insmod galcore.ko
```

二、替换动态库（可选，公版方案中已预置动态库）

VIPLite 版本动态库有：

```
├── libNBGlinker.so
└── libVIPhal.so
```

若无预置，使用adb push方式将动态库推到板端目录中，Android 系统可从android_sdk/vendor/aw/ai-sdk/viplite-android/v2.0中获取，推到/vendor/lib64或者其他目录下；Linux 系统可从tina_sdk/platform/allwinner/vision/ai-sdk/viplite-tina/lib/aarch64-none-linux-gnu/v2.0，推到/usr/lib或其他目录下。

Unified 版本动态库有（从android_sdk/vendor/aw/ai-sdk/unified-android/中获取）：

```
├── libCLC.so
├── libGAL.so
├── libGLSLC.so
├── libNNGPUBinary.so
├── libNNVXCBinary.so
├── libOpenVX.so
├── libOpenVX.so.1
├── libOpenVX.so.1.3.0
├── libOpenVXU.so
├── libOvx12VXCBinary.so
├── libOvxGPUVXCBinary.so
└── libovxlib.so
```

若推到其他目录下时，请配置环境变量（使用静态库编译例程时不需要，使用系统内置的动态库不需要）：

```
export LD_LIBRARY_PATH=/xxx/xxx
```

三、临时修改频率

```
#打开调试节点，若已有则无需重复执行
mount -t debugfs none /sys/kernel/debug
#查看可用频点
cat /sys/kernel/debug/viplite/clk_freq
supported frequencies:
    492000000 (Hz)
    852000000 (Hz) <---- current freq

#临时修改频率，可用以下命令
```

版权所有 © 珠海全志科技股份有限公司。保留一切权利

```
echo 492000000 > /sys/kernel/debug/viplite/clk_freq
```

版权所有 © 珠海全志科技股份有限公司。保留一切权利

3 测试步骤

说明

本文以 VIPLite 版本 NPU 驱动为例，使用 vpm_run 程序进行测试。vpm_run 程序使用方法请参考《NPU_ 模型部署 _ 开发指南》。公版方案中已预置 vpm_run 程序，若无预置参考《NPU_ 模型部署 _ 开发指南》的《模型运行工具 vpm_run 使用说明》一节进行编译。

1. 将测试用例以及模型和输入使用adb push传入板端；
2. 配置sample.txt指定模型与输入；

```
[network]
./network_binary.nb
[input]
./input.dat
```

3. 运行测试用例：

```
chmod +x vpm_run
##运行例程 -l 100：推理100次
./vpm_run -s sample.txt -l 100
```

板端已预置 vpm_run 程序，Andoird 系统在/vendor/bin/vpm_run 目录，Linux 系统在/usr/bin/vpm_run，可直接使用。

4. 输出结果，分析 LOG 得出结论。

```
a733-pro3:/vendor/bin # vpm_run -s /vendor/etc/input_data/sample.txt -l 100
loop_count=100, device_index=0, core_index=-1, file_name=/vendor/etc/input_data/sample.txt, time_out=0x0, bypass
        =1
enable_npd=0, preload=0
show_top50, save_txt=0
init vip lite, driver version=0x00020003...
VIPLite driver software version 2.0.3.0-AW-2024-05-29

vip lite init OK.

cid=0x1000003b, device_count=1
  device[0] core_count=1
config file read network count=1
init test resources, task_count: 1 ...
create/prepare networks ...
task i=0, binary name: /vendor/etc/models/network_binary.nb
nbg name=/vendor/etc/models/network_binary.nb
create network 0: 1967 us.
```

版权所有 © 珠海全志科技股份有限公司。保留一切权利

```
input 0 dim 224 224 3 1, data_format=2, quant_format=2, name=input[0], scale=0.003922, zero_point=0
ouput 0 dim 2 1 0 0, data_format=2, name=uid_1_out_0, scale=0.001625, zero_point=128
 memory pool size=1092352byte
network core count=1
prepare network 0: 1116 us.
golden file count=0
input 0 name: /vendor/etc/input_data/input_0.dat
read input and golden 0: 549 us.
task: 0, loop count: 1
start to run network=/vendor/etc/models/network_binary.nb
run time for this network 0: 5476 us.
run network done...
#第1次推理时间
profile inference time=4703us, cycle=3840288
...
...

task: 0, loop count: 100
start to run network=/vendor/etc/models/network_binary.nb
run time for this network 0: 4123 us.
run network done...
#第100次推理时间
profile inference time=3263us, cycle=2418713
#分析log，可知平均推理时间为4703us
task 0, profile avg inference time=4703us, cycle=3840288
destroy test resource task_count=1
vpm run ret=0
```

Linux 系统对应命令可参考以下：

```
cd /etc/npu/vpm_run
vpm_run -s sample.txt -l 100
```

版权所有 © 珠海全志科技股份有限公司。保留一切权利

4 公开模型的性能测试

测试各模型的帧率（推理 1000 次）、带宽占用等情况，基于 2400Mhz 的 LPDDR5 和 852MHz 频率的 NPU 模块进行测试。

模型名称	输入分辨率	量化精度	FPS(852)	ReadBW(MB)	WriteBW(MB)
yolov3	3x416x416	uint8	23.46	104.723	37.657
yolov4-tiny	3x416x416	uint8	182.68	8.281	4.208
yolov5s-sim	3x640x640	uint8	51.66	59.4	25.382
yolov5m	3x608x608	uint8	24.46	101.069	58.291
yolov7	3x640x640	uint8	10.81	334.935	78.479
yolov9c	3x640x640	uint8	8.79	396.905	93.017
ssd_mobilenet_v1	300x300x3	uint8	283.61	7.154	1.819
alexnet	3x227x227	uint8	62.34	45.96	0.339
inception_v3	299x299x3	uint8	51.12	34.933	4.998
inception_v4	299x299x3	uint8	23.44	83.804	17.09
mobilenetv1_1.0	224x224x3	uint8	636.54	3.701	0.508
mobilenetv2-12	3x224x224	uint8	581.4	5.513	1.736
resnet50v2	3x224x224	uint8	112.32	32.568	8.583
fairmot_crowdhuman	3x608x1088	uint8	5.28	342.468	463.531
yolact_resnet50	3x550x550	uint8	9.09	339.771	134.723
openpose	3x184x184	uint8	34.16	57.8	3.776
alphaPose_hard68	3x256x192	uint8	7.52	47.416	13.913
handpose_x	3x256x256	uint8	193.09	18.555	9.544
PFLD98-sim	3x112x112	uint8	1308.9	1.259	0.058
Swin_MLP	3x224x224	uint8	31.12	69.332	86.487
Swin_Transformer	3x224x224	uint8	17.25	136.199	130.089
Swin_Transformer_v2	3x224x224	uint8	8.16	236.75	205.645
SimpleViT-sim	3x256x256	uint8	40.7	85.492	19.877
T2TViT-sim	3x256x256	uint8	1.9	1319.091	179.627

版权所有 © 珠海全志科技股份有限公司。保留一切权利

5 公开模型的资源情况

以下为测试的公开模型的资源占用情况，内存占用统计使用了 VmHWM 值：

模型名称	输入分辨率	量化精度	模型大小（MB）	内存占用（MB）
yolov3	3x416x416	uint8	37.17	38.38
yolov4-tiny	3x416x416	uint8	3.83	4.5
yolov5s-sim	3x640x640	uint8	5.32	6.38
yolov5m	3x608x608	uint8	14.78	17.5
yolov7	3x640x640	uint8	21.93	23.75
yolov9c	3x640x640	uint8	18.14	22
ssd_mobilenet_v1	300x300x3	uint8	4.53	5.12
alexnet	3x227x227	uint8	44.98	45.25
inception_v3	299x299x3	uint8	16.42	16.5
inception_v4	299x299x3	uint8	28.84	29.12
mobilenetv1_1.0	224x224x3	uint8	3.04	3.38
mobilenetv2-12	3x224x224	uint8	3.01	3.38
resnet50v2	3x224x224	uint8	16.8	17.25
fairmot_crowdhuman	3x608x1088	uint8	15.31	17.88
yolact_resnet50	3x550x550	uint8	24.35	26.12
openpose	3x184x184	uint8	31.77	32
alphaPose_hard68	3x256x192	uint8	20.4	20.62
handpose_x	3x256x256	uint8	2.64	2.75
PFLD98-sim	3x112x112	uint8	1.07	1.25
Swin_MLP	3x224x224	uint8	17.36	19
Swin_Transformer	3x224x224	uint8	27.92	31.38
Swin_Transformer_v2	3x224x224	uint8	42.21	52.12
SimpleViT-sim	3x256x256	uint8	55.04	55.62
T2TViT-sim	3x256x256	uint8	28.92	35.5

版权所有 © 珠海全志科技股份有限公司。保留一切权利

6 附录

公开模型获取途径如下：

分类	模型	获取途径
目标检测	yolov3	cfg、weights
目标检测	yolov5s-sim	下载
目标检测	yolov9c	下载
目标检测	ssd_mobilenet_v1	下载
分类	alexnet	下载，权重
分类	resnet50v2	下载
分类	mobilenetv1_1.0	下载
分类	mobilenetv2-12	下载
分类	inception_v3	下载
分类	inception_v4	下载
多目标跟踪	fairmot_crowdhuman	下载
实例分割	yolact_resnet50	下载
人体姿态估计	openpose	下载
人体姿态估计	alphaPose_hard68	下载
手部关键点	handpose_x	下载
人脸关键点	PFLD98-sim	下载
Transformer 骨干	Swin_Transformer	下载
Transformer 骨干	Swin_Transformer_v2	下载
MLP	Swin_MLP	下载
Transformer 骨干	T2TViT-sim	下载
Transformer 骨干	SimpleViT-sim	下载

版权所有 © 珠海全志科技股份有限公司。保留一切权利

著作权声明

版权所有 ©2025 珠海全志科技股份有限公司。保留一切权利。

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。