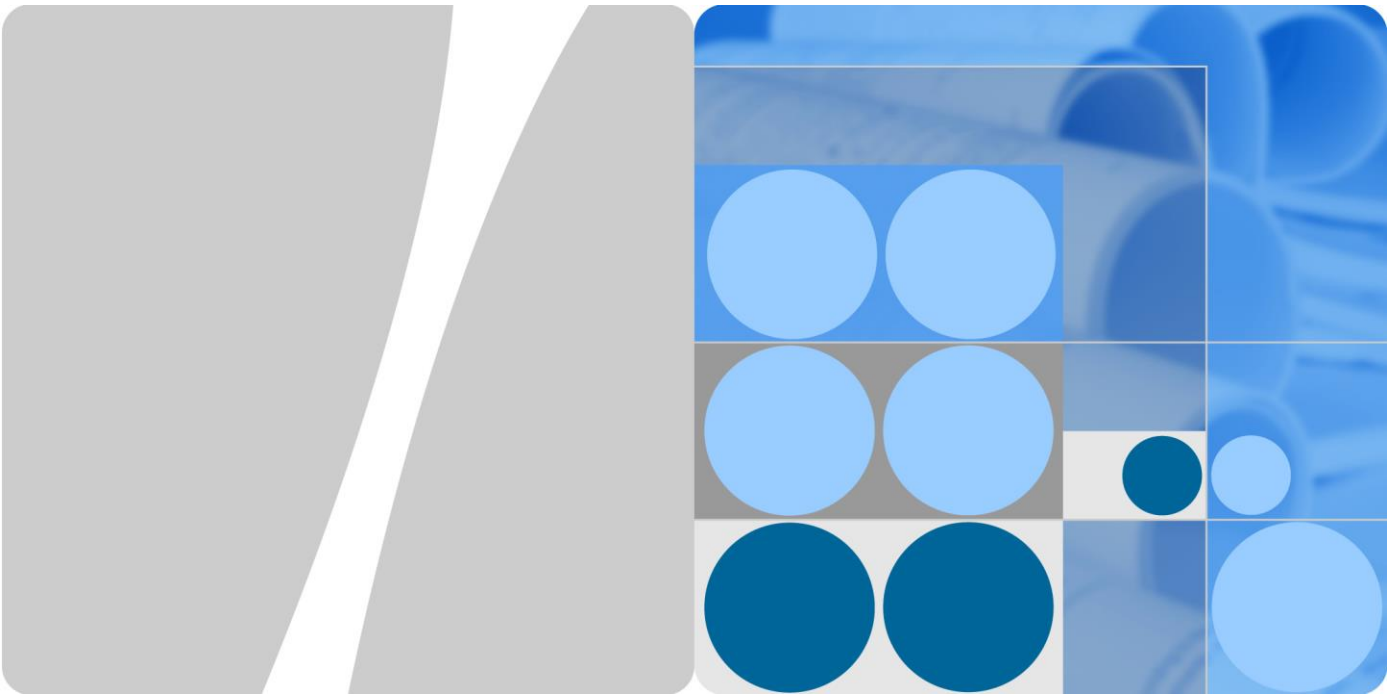


HiAI DDK V320

快速入门

文档版本 04

发布日期 2020-02-28

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。

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。

法律声明

本文所描述内容可能包含但不限于对非华为或开源软件的介绍或引用，使用它们时请遵循对方的版权要求。

注意

您购买的产品、服务或特性等应受华为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，华为公司对本文档内容不做任何明示或暗示的声明或保证。

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。

华为 HiAI 申请方式

发送申请邮件到邮箱：developer@huawei.com

邮件名称：HUAWEI HiAI+公司名称+产品名称

邮件正文：合作公司+联系人+联系方式+联系邮箱

我们将在收到邮件的 5 个工作日内邮件给您反馈结果，请您注意查收。

官网地址 https://developer.huawei.com/consumer/cn/

前言

概述

本文提供对华为 HiAI DDK V320 的整体介绍以及对集成操作流程的简要说明。

本文与以下文档配套使用：

文档名称	描述说明
华为 HiAI_DDK_V320_版本说明书	介绍 HiAI DDK_V320 和之前版本的变化以及特性。
华为 HiAI_DDK_V320_FAQ	介绍 HiAI DDK 相关 FAQ。
华为 HiAI_DDK_V320_IR 模型构建使用说明书	介绍模型构建的方式以及接口，支持除 TensorFlow 和 Caffe 外的框架算子级集成对接。
华为 HiAI_DDK_V320_快速入门	介绍 HiAI DDK 内容。
华为 HiAI_DDK_V320_模型推理集成指导	介绍模型集成与编译方式及集成接口。
华为 HiAI_DDK_V320_OMG 工具使用说明	介绍离线转换工具 OMG 的使用。
华为 HiAI_DDK_V320_轻量化工具使用说明书	介绍相关轻量化工具的使用。
华为 HiAI_DDK_V320_算子规格说明	介绍 HiAI DDK V320 支持算子的边界条件。
华为 HiAI_DDK_V320_缩略语	介绍 HiAI DDK 中缩略语和专有名词。
华为 HiAI_DDK_V320_系统调试工具使用说明书	介绍系统调试工具的使用。

修改记录

日期	修订版本	修改描述
2020-02-28	04	新增系统调试工具和支持通用 ARM 处理器
2019-12-31	03	新增 V320 版本内容
2019-11-18	02	增加 Demo 中开启 AIPP 功能的提示
2019-09-06	01	新增 V310 版本内容

目 录

插图目录

表格目录

1 HiAI 简介

HiAI 是面向移动终端的 AI 计算平台。HiAI DDK（Device Development Kit）是对第三方开发者开放的 HiAI 资源包。

HiAI API 是移动计算平台中的人工智能计算库，该计算库面向人工智能应用程序开发人员，让开发者便捷高效地编写在移动设备上运行的人工智能应用程序。

HiAI API 以统一的二进制文件发布，主要作用是通过 HiAI 异构计算平台来加速神经网络的计算，当前仅支持在 Kirin SoC(System on a Chip)上运行。

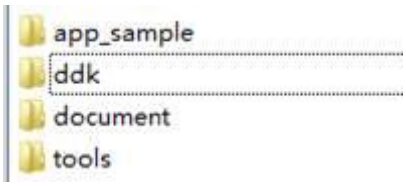
HiAI API 集成在使用 Kirin SoC 芯片的 android 系统上，开发者可以在集成环境中运行神经网络模型，调用 HiAI API 进行加速计算。

2 DDK 包说明

2.1 概述

DDK(Device Development Kit)是 HiAI 开放的开发包，一个完整的 HiAI DDK 包如图 2-1 所示。

图2-1 DDK 包内容



- app_sample 为 Android demo app 的源码。
- ddk 为 HiAI 开放的 SDK。
- document 为开发参考文档。
- tools 为 OMG 离线转换工具和轻量化工具。

说明

app_sample 目录中的 IR_model_demo 只能运行在华为自研 NPU 手机。

2.2 sample

以 sample\inference_npu_demo\Demo_Source_Code.中 SqueezeNet 分类网络模型为例（下载地址：https://github.com/DeepScale/SqueezeNet/tree/master/SqueezeNet_v1.1）提供了从输入前处理到模型加载、模型前向计算、前向计算结果后处理、模型卸载、时间统计等一系列 sample code 示例，同时提供了同步，异步方式的 sample code。V320 中提供的 AIPP 功能，可以进行输入预处理，仅 Kirin990 支持，sample code 中加载 AIPP 模型的代码，默认是注释的，可以自行开启，位置在 Demo_Soure_Code\app\src\main\java\com\huawei\hiaidemo\view\MainActivity.java 文件的 MainActivity 类的 initModels()初始化模型函数中。

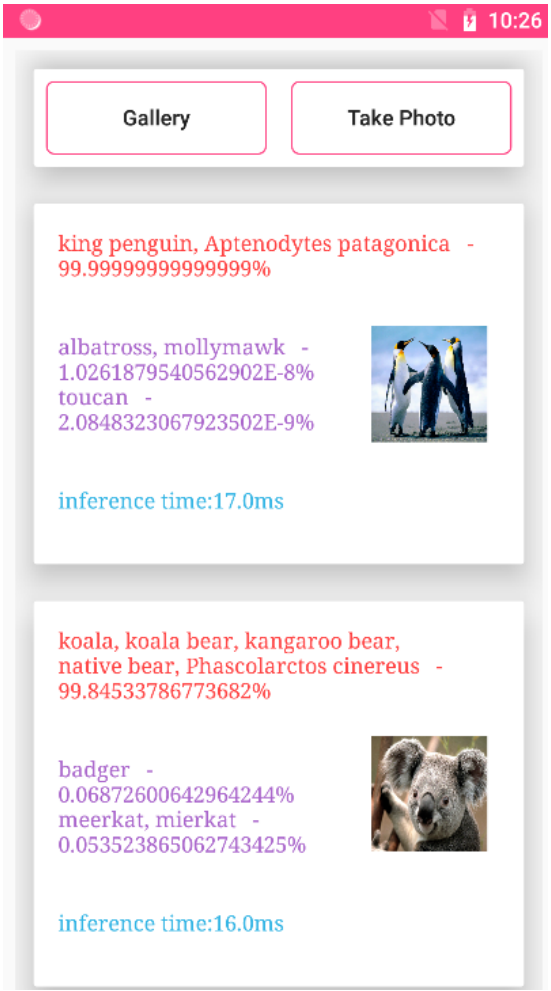
```
ModelInfo model_1 = new ModelInfo();
model_1.setModelSaveDir(path);
model_1.setUseAIPP(true);
model_1.setOfflineModel("hiai.om");
model_1.setOfflineModelName("hiai");
model_1.setOnlineModelLabel("labels_caffe.txt");
demoModelList.add(model_1);
```

说明

使用 Android Studio 2.2 及以上版本（请参考 Google Android Developers 网站：https://developer.android.com/studio/index.html ）。

导入 sample 源码并运行，可以看到 App 支持从图库中选择图片或者拍照进行分类。App 运行效果图如图 2-2。

图2-2 DDK App 运行效果



2.3 ddk

ddk 中的 ai_ddk_lib 文件夹包含下面两部分：

- 模型推理：依赖库和相关头文件。

目录文件	描述说明
ai_ddk_lib\lib64\libhiai.so	DDK 使用 NPU 进行模型推理需要依赖的动态库
ai_ddk_lib\lib64\libhcl.so	DDK 使用 NPU 进行模型推理需要依赖的动态库
ai_ddk_lib\lib64\libcpucl.so	DDK 使用 CPU 进行模型推理需要依赖的动态库（可选）
ai_ddk_lib\include\HiAiModelManagerService.h	DDK 对外提供 C++接口头文件
ai_ddk_lib\include\HiAiModelManagerType.h	DDK 对外提供 C++类型定义头文件
ai_ddk_lib\include\HiAiAippPara.h	DDK 对外提供 C++ AIPP 接口定义头文件（可选）
ai_ddk_lib\include\hiai_types.h	DDK 对外提供 C++类型定义头文件
ai_ddk_lib\include\native_handle.h	DDK 依赖头文件

- 模型构建：依赖库和相关头文件。

目录文件	描述说明
ai_ddk_lib\lib64\libhiai_ir.so	IR 算子定义 & 图构建依赖库
ai_ddk_lib\lib64\libhiai_ir_build.so	IR 模型编译依赖库
ai_ddk_lib\include\hiai_ir_build.h	DDK IR API 构建、算子定义、模型编译等接口头文件

2.4 document

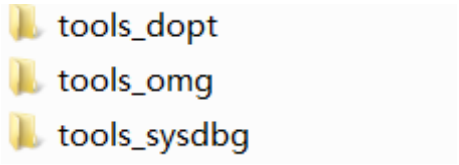
该文件夹包含以下文档：

文档名称	描述说明
华为 HiAI_DDK_V320_版本说明书	介绍 HiAI DDK_V320 和之前版本的变化以及特性。
华为	介绍 HiAI DDK 相关 FAQ。

文档名称	描述说明
HiAI_DDK_V320_FAQ	
华为 HiAI_DDK_V320_IR 模型构建使用说明书	介绍模型构建的方式以及接口，支持除 TensorFlow 和 Caffe 外的框架算子级集成对接。
华为 HiAI_DDK_V320_快速入门	介绍 HiAI DDK 内容。
华为 HiAI_DDK_V320_模型推理集成指导	介绍模型集成与编译方式及集成接口。
华为 HiAI_DDK_V320_算子规格说明	介绍 HiAI DDK V320 支持算子的边界条件。
华为 HiAI_DDK_V320_缩略语	介绍 HiAI DDK 中缩略语和专有名词。
华为 HiAI_DDK_V320_轻量化工具使用说明书	轻量化工具使用的指导说明书
华为 HiAI_DDK_V320_OMG 工具使用说明	OMG 工具使用的指导说明书
华为 HiAI_DDK_V320_系统调试工具使用说明书	系统调试工具使用的指导说明书

2.5 tools

tools 目录包含以下三个目录，如下图：



- tools_dopt 为轻量化工具操作指导及 demo 样例
- tools_omg 为 Linux 平台下 Caffe 和 TensorFlow OMG 离线转换工具以及离线构建 demo 样例
- tools_sysdbg 为 Android 平台下，系统调试工具

各目录具体结构如下：

表2-1 tools_dopt 目录

目录文件	描述说明

目录文件	描述说明
tools\tools_dopt\caffe	对应 Caffe 框架重训练使用的 so 以及源码
tools\tools_dopt\tensorflow	对应 TensorFlow 框架重训练使用的 so
tools\tools_dopt\dopt_trans_tools	重训练后转换模型的工具
tools\tools_dopt\demo	提供 Caffe 和 TensorFlow 的样例模型
tools\tools_dopt\config	为用户使用的框架信息的配置脚本，例如 Caffe 的路径
tools\tools_dopt\env	轻量化工具 Docker 环境配置文件

表2-2 tools_omg 目录

目录文件	描述说明
tools\tools_omg\omg	离线模型转换工具
tools\tools_omg\v300	V300 能力包，当前 OMG 指定版本为 V300 时会使用此目录的功能
tools\tools_omg\v310	V310 能力包，当前 OMG 指定版本为 V310 时会使用此目录的功能
tools\tools_omg\v320	V320 能力包，当前 OMG 指定版本为 V320 时会使用此目录的功能
tools\tools_omg\IR	IR 能力包，当前 OMG 指定版本为 IR 时会使用此目录的功能
tools\tools_omg\sample	AIPP 转换，8bit 量化转换的示例模型以及配置文件

表2-3 tools_sysdbg 目录

目录文件	描述说明
tools\tools_sysdbg\data_proc_tool	性能数据处理工具(生成*.csv)
tools\tools_sysdbg\model_run_tool	性能数据生成工具
tools\tools_sysdbg*.so	工具运行依赖库

3 DDK 版本与 Kirin 平台映射关系

DDK 版本	典型终端型号	Kirin 平台	HiAI Version	支持算子数
V150	P20 P20 Pro Mate RS Honor 10	Kirin970	-	90
	Nova 3 Honor play Honor Note10			
V200	Mate20 Mate20 Pro	Kirin980	-	150
V300	Nova 5 Nova 5z Nova 5i pro Honor 9X Honor 20s	Kirin810	100.300.xxx.xxx	178
V310	Mate 30	Kirin990	100.310.xxx.xxx	223
V320	P40	Kirin990 Kirin820 Kirin985	100.320.xxx.xxx	306

4 集成操作流程

集成操作流程是指导用户将原始模型通过离线模型转换工具生成 OM 模型，再通过模型推理集成生成 APK，将 APK 在 Kirin SOC 上运行，神经网络加速。

4.1 环境准备

- 使用 Ubuntu 16.04、macOS 安装应用开发环境 Android Studio。

  Android Studio 下载地址：https://developer.android.com/studio/index.html
- 使用 NDK 进行 Native 代码编译，推荐 NDK r14b 及其以上版本。另外，可以使用 CMake 进行 Native 代码编译。

  NDK 下载地址：https://developer.android.com/ndk/downloads/index.html
- 使用 Ubuntu 16.04 64 位 运行 tools_omg 模型转换工具。

  Linux 各镜像下载地址: http://mirrors.ustc.edu.cn/ubuntu-releases/16.04/
- 准备训练好的 Caffe 或 TensorFlow 模型。
- 准备搭载 Kirin 平台的设备用于测试 APP，具体平台与终端型号参见“3 DDK 版本与 Kirin 平台映射关系”。

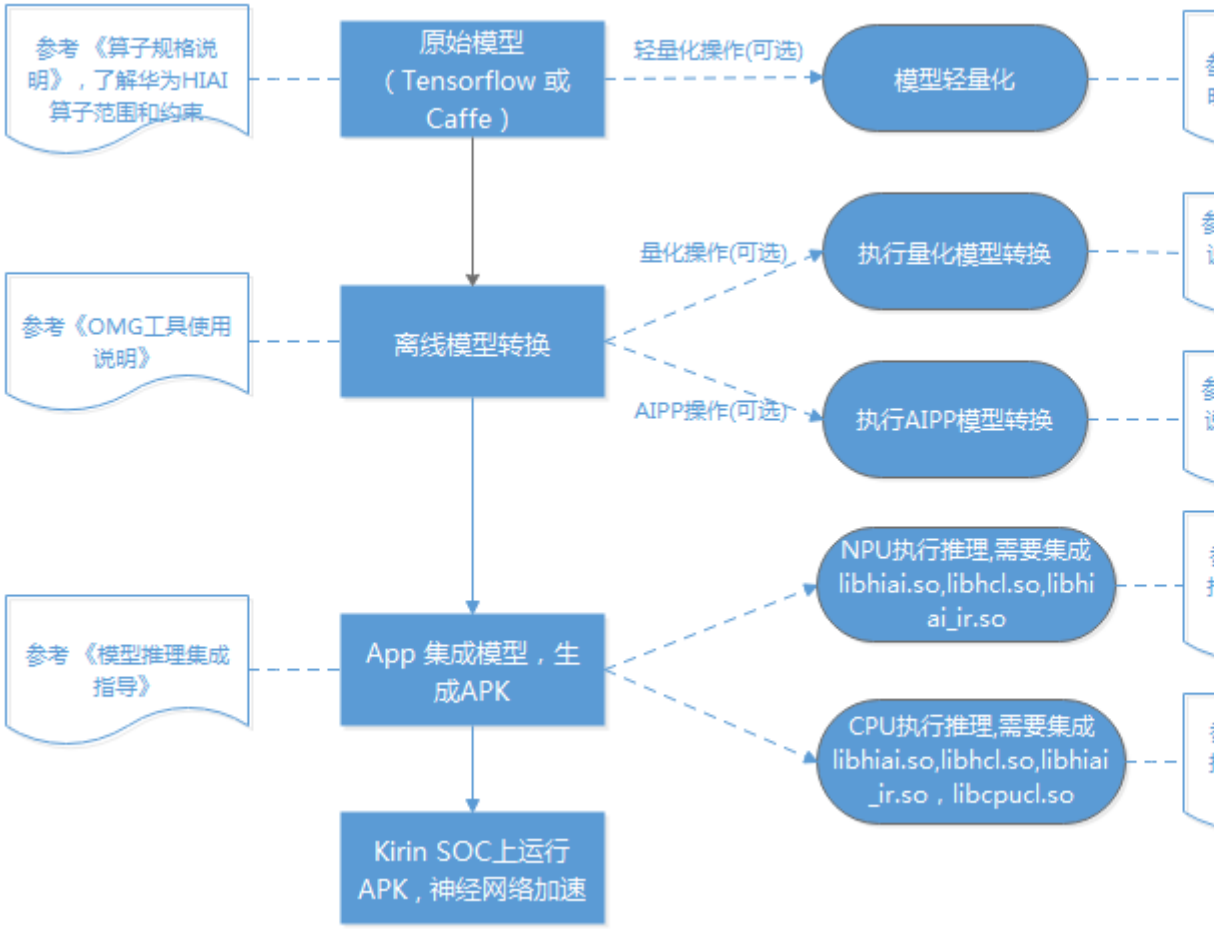
4.2 流程说明

下图介绍了 App 使用 HiAI DDK V320 的集成流程，其中 AIPP（Artificial Intelligence Pre-processing）、量化为可选步骤。

说明

下图中的参考文档请根据实际应用版本进行参考。

图4-1 NPU/CPU 集成操作流程图



原始模型轻量化

针对原始模型（Tensorflow 或 Caffe）进行深度的模型优化，可以帮助用户自动完成模型轻量化，达到减少模型体积以及加快模型推理速度的目的。目前支持无训练模式和重训练模式。轻量化操作可参考《轻量化工具使用说明书》。

离线模型转换

离线模型转换需要将 Caffe 或者 TensorFlow 模型转换为 HiAI 平台支持的模型格式，并可以按需进行 AIPP 操作、量化操作，使用场景及方法如下：

- AIPP 操作

  AIPP 用于在硬件上完成图像预处理，包括改变图像尺寸、色域转换（转换图像格式）、减均值/乘系数（改变图像像素），运用后可避免重新训练匹配推理计算平台需要的数据格式，仅仅通过 AIPP 参数配置或者在软件层面上调用 AIPP 接口即可完成适配，同时由于硬件专用，可以获得较好的推理性能收益，具体操作可参考《华为 HiAI_DDK_V320_OMG 工具使用说明》中 AIPP 模型转换以及配置的操作指导。
- 量化操作

量化是一种可以把 fp32 模型转化为低 bit 模型的操作，以节约网络存储空间、降低传输时延以及提高运算执行效率，量化操作可参考《华为 HiAI_DDK_V320_OMG 工具使用说明》中量化模型转换的操作指导。

APP 集成

APP 集成流程包含模型预处理、加载模型、运行模型、模型后处理。

- NPU 场景下，APP 需要在模型预处理过中集成 libhiai.so、libhcl.so、libhiai_ir.so，编译 APK 后可在 NPU 上执行推理，参考《华为 HiAI_DDK_V320_模型推理集成指导》。
- CPU 场景下，APP 需要在模型预处理过中集成 libhiai.so、libhcl.so、libhiai_ir.so 以及 libcpucl.so，编译 APK 后可在 CPU 上执行推理，参考《华为 HiAI_DDK_V320_模型推理集成指导》。

5 支持的算子

请参见《华为 HiAI_DDK_V320_算子规格说明》。

6 DDK 数据安全说明

6.1 DDK 工作方式

移动端 DDK 需要在应用打包时，被加载在您的应用当中。DDK 会随着客户应用的启动开始进行加载。当用户关闭应用时，DDK 会随着客户应用的关闭而关闭，不会在后台做任何额外动作。

6.2 DDK 权限说明

DDK 不涉及权限申请。

6.3 DDK 收集数据

DDK 不采集任何数据，仅接受您的应用传递的数据。

6.4 DDK 数据安全保护

DDK 接收的数据仅在端侧处理，不涉及上报服务器。

