基于华为自研AI芯片的计算机视觉深度学习算法研究与移植

测试方案

浙江大学

## 1 概述

本项目将OpenPose和MS-G3D相结合，并在Atlas 200 DK上进行了移植与部署，构建了一个实时动作检测系统。该系统可通过外设摄像头捕获实时视频，送入部署在Atlas200DK上的深度学习模型进行推理计算后得到检测结果，将标出结果的视频通过Presenter Server在浏览器上播放。

本测试方案针对基于Atlas 200 DK开发板的超分辨率图像转换系统制定，开发板的DDK版本为1.3.0.0版本。本测试方案主要对系统进行功能测试以及性能测试，包括不同超分辨率图像算法在Atlas 200 DK上的推理时间、在Set5数据集上的PSNR，以及模型压缩前后的性能对比，进而评估该系统的质量与实用性。

## 2 系统环境

本部分介绍本系统的测试环境，开发环境和开发工具。

### 2.1 测试环境

测试环境即该系统的应用环境，该环境分为两部分，分别是Host端环境以及device端环境。Host端指与Atlas200DK相连的ubuntu系统的电脑，负责事先向Atlas200DK传输需要处理的图像，以及接收并查看超分辨率网络的生成结果。Device端即Atlas200DK开发板端，Device端主要负责图像的读取、预处理，以及超分辨率模型的运行等过程。

具体的应用环境配置如表2-1所示。

表2-1 应用环境配置

|  |  |  |  |
| --- | --- | --- | --- |
|  | 硬件 | 系统 | 环境要求 |
| 摄像头 | 摄像头 | ------- | --------- |
| Host端 | 电脑、RAM>4G | ubuntu-18.04-desktop-amd64 | 已配置交叉编译器 |
| Device端 | Atlas200DK | ubuntu-18.04-server-armd64 | 已配置交叉编译器 |

### 2.2 开发环境

本系统用到的开发工具如表2-2所示。

表2-2 系统开发工具

|  |  |  |
| --- | --- | --- |
| 工具名称 | 工具类别 | 作用 |
| TensorFlow/Pytorch | 深度学习框架 | 模型网络训练 |
| MindStudio | 集成开发环境 | 模型转换 |
| OpenCV | 计算机视觉库 | 图像读取、保存等处理 |
| Atlas200DK | 项目运行平台 | 系统运行和网络推理 |

## 3 功能测试

该系统能够从本地文件中读取低分辨率图像数据，送入超分辨率网络模型进行推理计算后得到放大倍数为3的高分辨率的图像，最后保存为本地文件以供查看。该系统的功能还包括能够选择不同的网络模型以及处理不同的图像形式（灰度、RGB）。本节我们使用Set5数据集（图3-1）中的图像进行功能测试，这里我们暂时不考虑系统性能，仅观测系统的功能性和完整性。

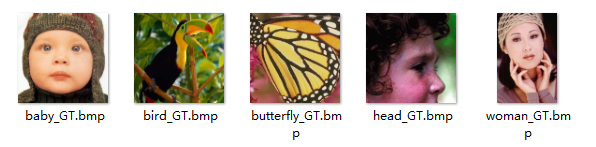


图3-1 Set5数据集

### 3.1 基础功能测试

1. 图像与模型上传

将sample-superresolution目录下提供的已经转换好的离线模型文件与Set5数据集上传至Atlas 200 DK开发板。

1. 应用执行与图像保存

以HwHiAiUser用户SSH登录到开发板侧，进入本项目对应的可执行文件所在路径，接着运行python脚本启动应用程序（请参考部署指导）。应用执行完毕后，在开发板侧会生成高分辨率图像以及用bicubic插值算法生成的对比图像。

如图3-2所示，输入命令之后系统提示“Success to deal file=…”表示已处理完成。另外可以看到当前目录下生成两张bmp图像，文件名中的“\_bicubic”和“\_srcnn”表示该图像是由bicubic或SRCNN算法生成的。

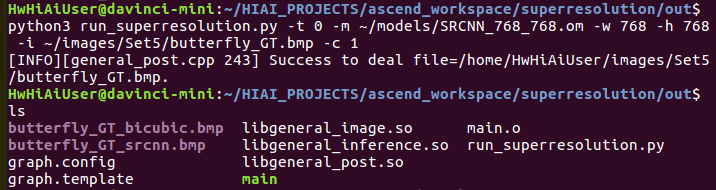


图3-2 应用执行

1. 图像结果下载

最后我们将图像下载回Ubuntu主机上进行查看。

LR bicubic SRCNN

图3-3运行结果

### 测试用例

## 4 性能测试

本节对本项目搭建的超分辨率图像系统在Atlas200DK平台上的性能进行测试，包括推理时间、PSNR，以及模型压缩前后的性能对比。

### 4.1 推理时间

本节以宽高为256×256的图像（Set5数据集中的butterfly\_GT.bmp）作为输入，记录了三种超分辨率网络在不同平台上的推演时间，如表4-1所示。其中，网络在CPU/GPU上的推演时间使用Caffe框架提供的计时方法time获得，而在Atlas 200 DK上，则是通过Profiling工具采集性能数据，以Model Statistic一栏中的Inference End Time - Inference Start Time作为推演时间。

表4-1 超分辨率算法模型的推演时间 (ms)

|  |  |  |  |
| --- | --- | --- | --- |
|  | CPU (Intel i7-7700) | GPU (GTX 1080 Ti) | Atlas 200 DK |
| SRCNN | 1336.18 | 8.37 | 20.69 |
| FSRCNN | 208.34 | 2.70 | 4.96 |
| ESPCN | 198.23 | 0.87 | 2.89 |

注：

输入为256×256图像，输出为768×768图像。

Atlas 200 DK使用的版本为1.3.T34.B891。

以上时间仅为网络推演耗时。对于SRCNN，推演时间不包括前处理bicubic的耗时；对于ESPCN，不包括后处理subpixel的耗时。

另测得subpixel操作在Atlas200DK后处理模块中耗时约40ms。

从表4-1的数据结果可以看出，本项目使用的三种超分辨率网络在Atlas 200 DK上的运行速度已远超其在Intel i7-7700 CPU上的速度，并且能满足指标要求，但相比于GTX 1080 Ti GPU还是存在一些差距。

### 4.4 测试用例