# 工大标志

# GPU计算实验报告

## 实验三 基于昇腾CANN的目标检测应用（AscendCL接口）

学院：计算学部

姓名：钱泽凯

学号：1190202011

### 一、实验预习（10分）

1. 注册华为云账号：https://www.huaweicloud.com/
2. 课程内容预习：

https://education.huaweicloud.com/courses/course-v1:HuaweiX+CBUCNXA024+Self-paced/about

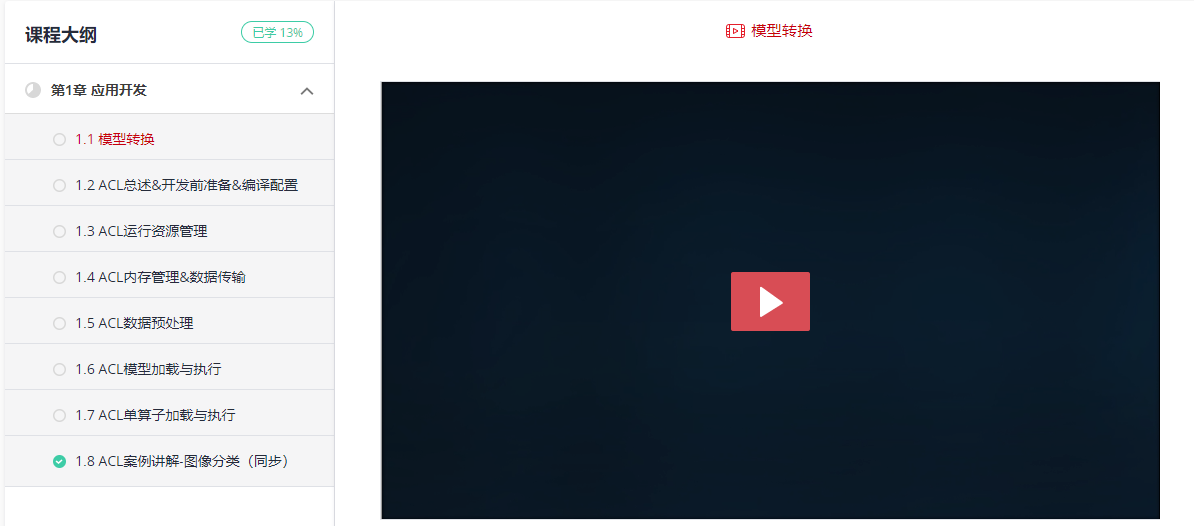


图1 昇腾学院发布的ASCENDCL课程

https://www.bilibili.com/video/BV16K4y1976L?spm\_id\_from=333.999.0.0

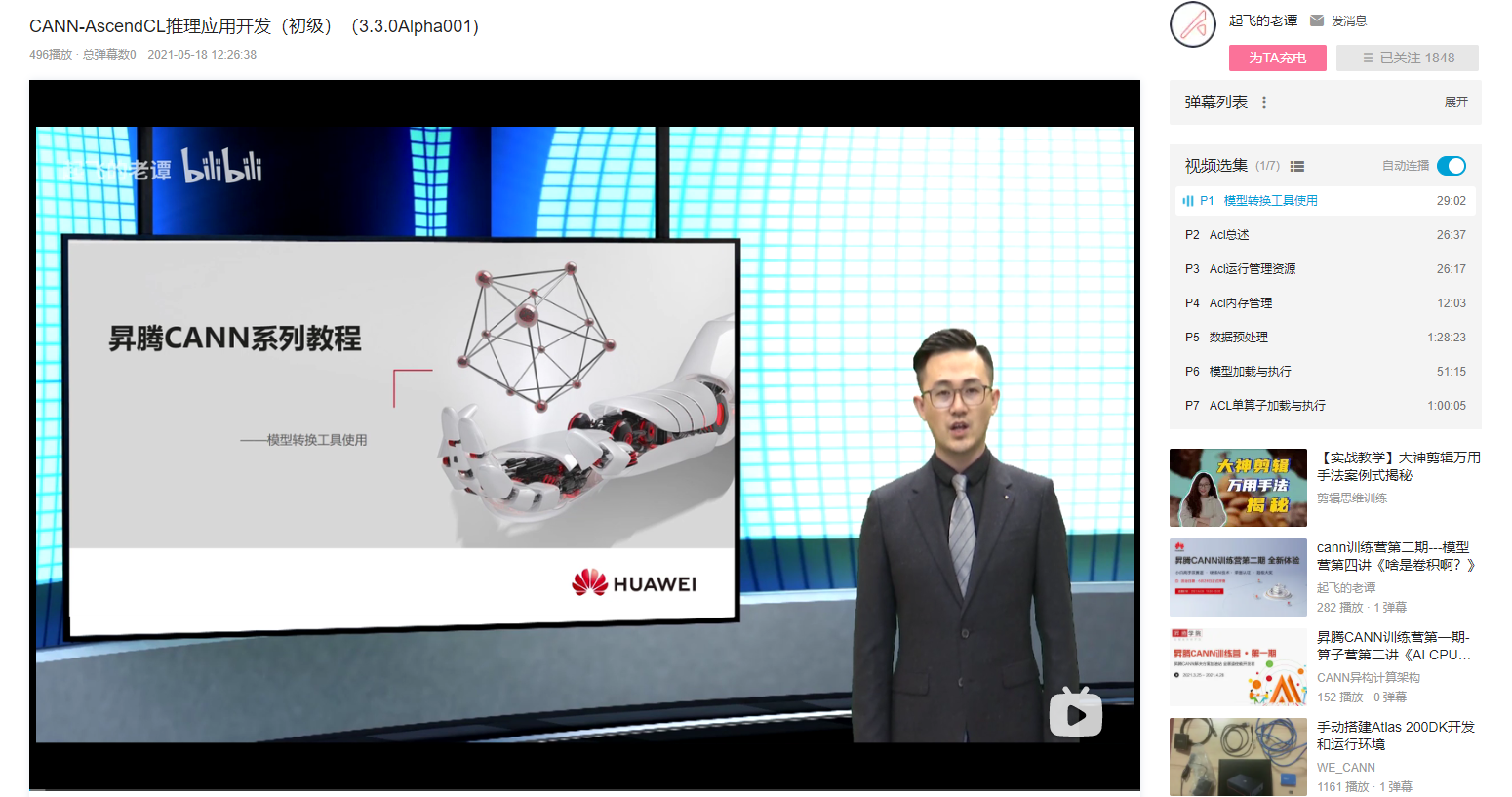


图2 华为谭老师bilibili发布的视频

1. 问题（10分）：

（回答请使用红色文字）

1. 在昇腾310上执行推理，模型格式必须是om，需要使用ATC工具进行模型转换，目前支持哪些深度学习框架的模型转换？ABC

A、Caffe B、TensorFlow C、MindSpore D、PyTorch

答：支持原始框架类型为Caffe、TensorFlow、MindSpore、ONNX的模型转换，其中ONNX可以看做是一个中间模型，如果其他模型可以转换为ONNX模型那么也可以在使用ATC进行模型转换，例如pytorch模型转换为ONNX模型，ONNX模型再转换为.om文件。

1. 模型转换时设置静态AIPP参数，模型生成后，AIPP参数被保存在om模型中，其功能主要有什么？ABC

A、图像裁切对齐 B、色域转换 C、归一化 D、图片解码与缩放

答：用于在AI Core上完成图像预处理，包括色域转换（转换图像格式）、图像归一化（减均值/乘系数）和抠图（指定抠图起始点，抠出神经网络需要的大小的图片）。

1. 下面那些选项是AscendCL能够完成的？ABDE
2. 运行资源管理（Device管理、Context管理、Stream管理）
3. 内存管理
4. 数据预处理DVPP
5. 模型加载与执行
6. 算子加载与执行

答：AscendCL（Ascend Computing Language）提供Device管理、Context管理、Stream管理、内存管理、模型加载与执行、算子加载与执行、媒体数据处理等C语言API库供用户开发深度神经网络应用，用于实现目标识别、图像分类等功能。

1. DVPP用于数据预处理，能够对图像解码、缩放，输出YUV420sp格式的图片，解码对输出图片的宽高有对齐要求，宽要求128对齐，高16对齐，缩放也对宽高有对齐要求，具体为：

答：要求宽16，高2对齐

宽：（16）对齐

高：（2）对齐

1. ASCENDCL中推理所需的输入输出数据，是通过一种特定的数据结构来组织的，称为Dataset，所有的输入组成了一个Dataset，所有的输出组成了一个Dataset，但对于很多模型来说，输入不止一个，那么所有的输入集合叫Dataset，其中的每一个输入叫什么？

答：databuffer

## 二、实验目标

1. 了解华为昇腾全栈开发工具MindStudio及其离线模型转换功能；
2. 了解如何使用ASCENDCL开发基于华为昇腾处理器的神经网络推理应用。

## 三、实验内容

完成链接中的实验：

https://lab.huaweicloud.com/testdetail.html?testId=458&ticket=ST-947652-UdMQSjr2EsqynL7E0Qu4Dkqz-sso



图3 基于昇腾AI处理器的目标检测应用（ASCENDCL）

参考文档：

https://support.huaweicloud.com/aclcppdevg-cann504alpha1infer/aclcppdevg\_01\_0001.html



图4 AscendCL官方文档

## 四、实验过程（80分）

1. 环境准备（10分）

完成弹性云服务器ECS创建，获取公网IP，如XX.XX.XX.XX，然后通过终端工具登录云服务器（密码为云服务器创建时自己设置的密码）：

ssh root@XX.XX.XX.XX

并且输入下列命令获取云服务器上昇腾芯片的信息并截图：

npu-smi info

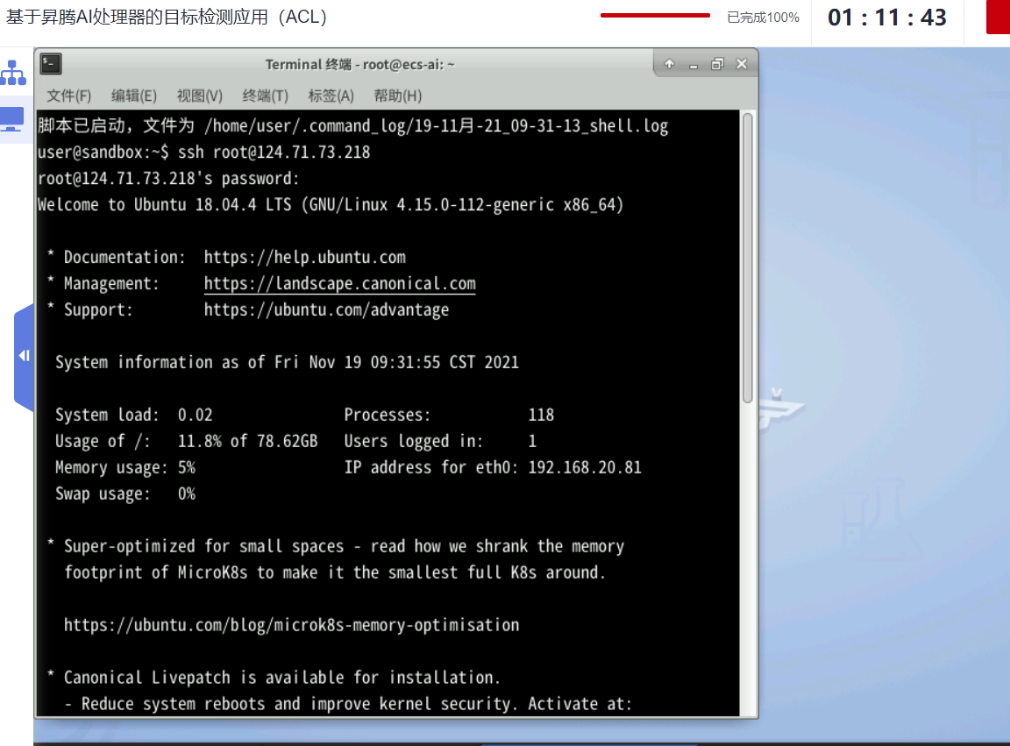


图1 登录云服务器

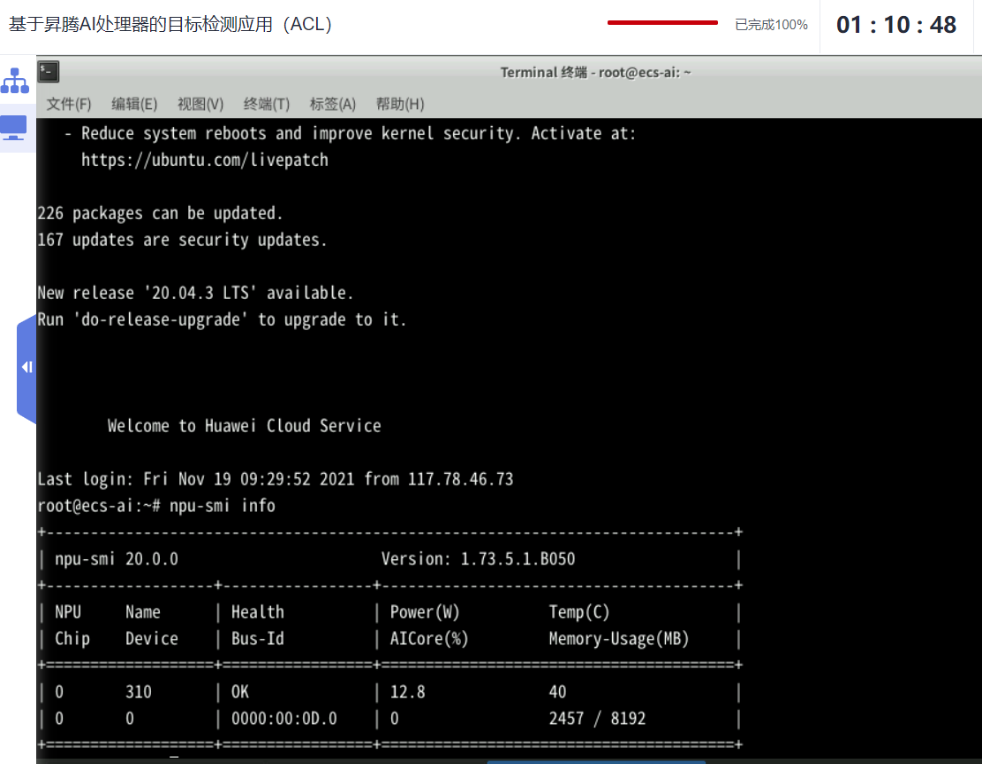


图2 查看昇腾芯片信息

2. 模型转换（30分）

对于已经转换成功的.om模型文件，可以在MindStudio界面呈现其网络拓扑结构，并可以查看模型所使用的算子：

https://support.huaweicloud.com/usermanual-mindstudio302/atlasms\_02\_0060.html

依次单击MindStudio菜单栏“Ascend > Model Visualizer”，选择转换成功的.om模型，将om模型可视化界面截图。

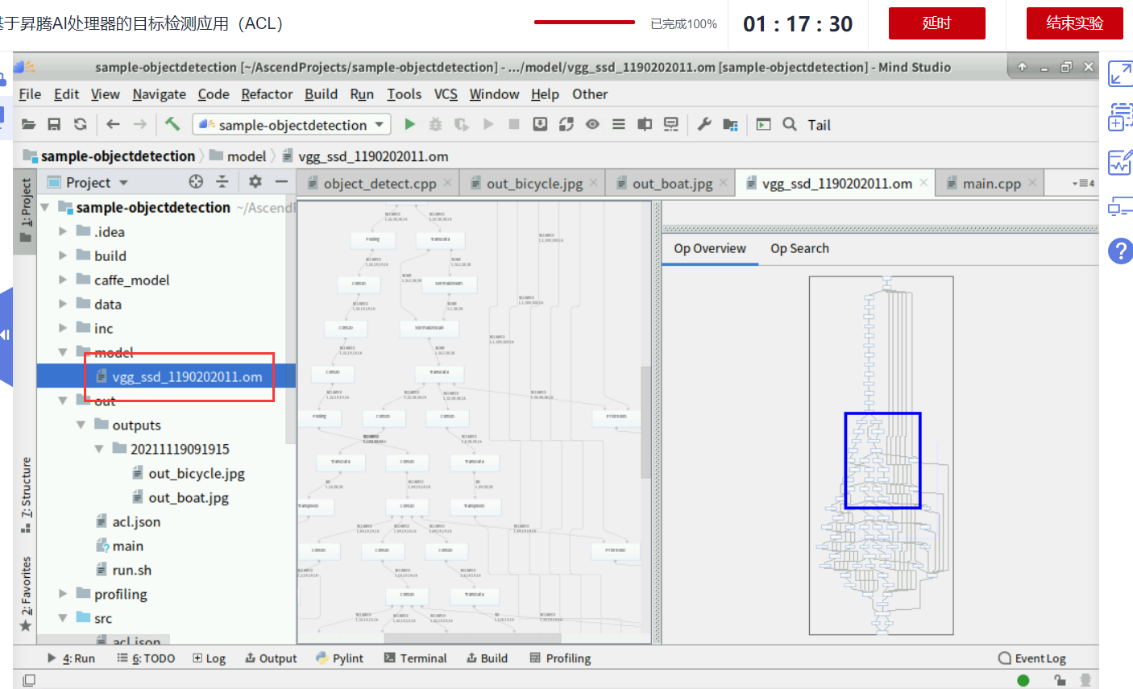
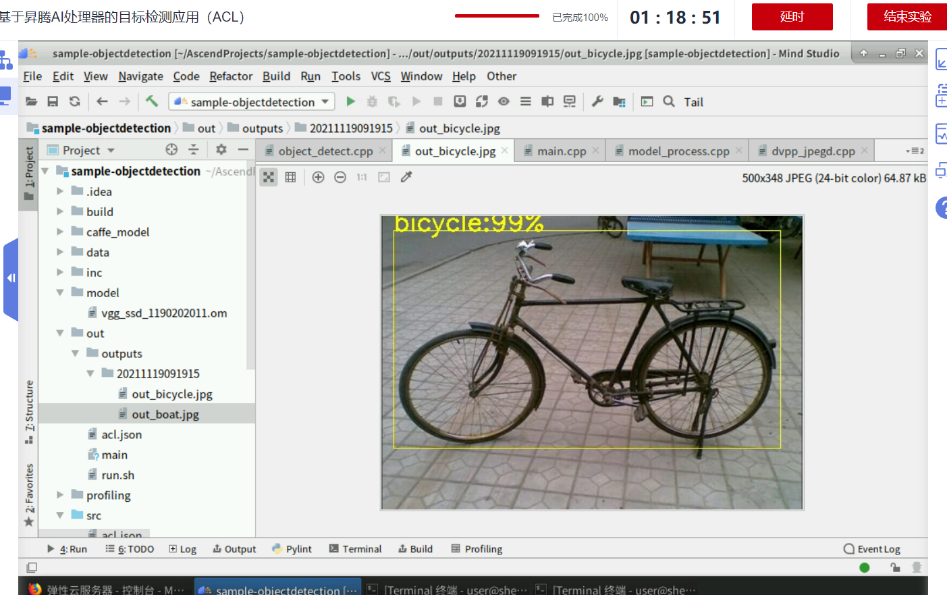


图3 om可视化界面，如图是以学号命名

3. 编译运行（30分）

按实验手册将代码补全，并编译运行，MindStudio会将工程发送到云服务器，完成推理后再将结果回传到沙箱环境，请将两个测试用例的推理结果截图。



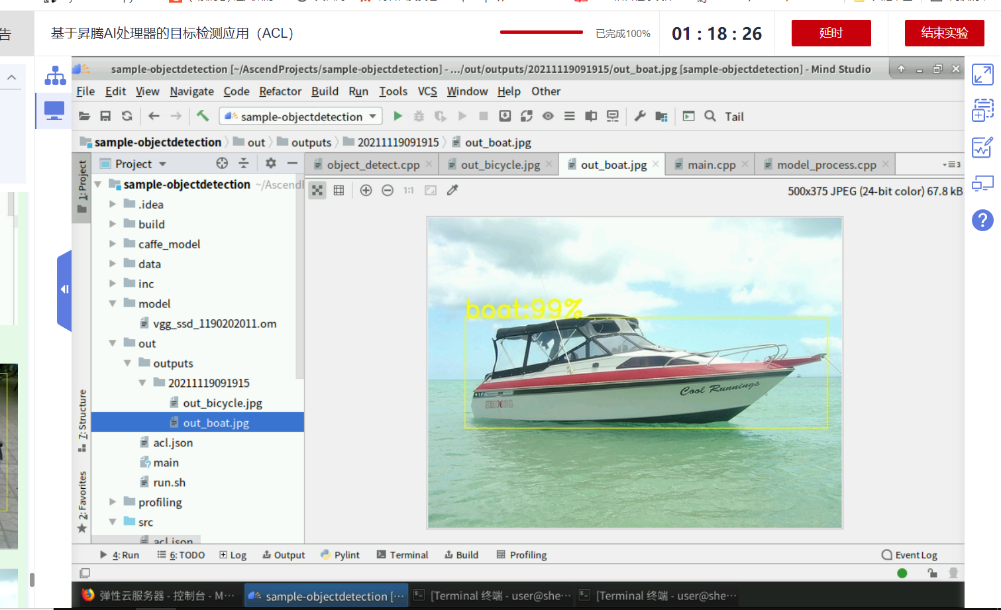


图4 模型运行结果

4. 模型性能（10分）

运行Profiling 查看推理性能，截图并指出模型从文件加载进内存耗时以及两张图片推理耗时。

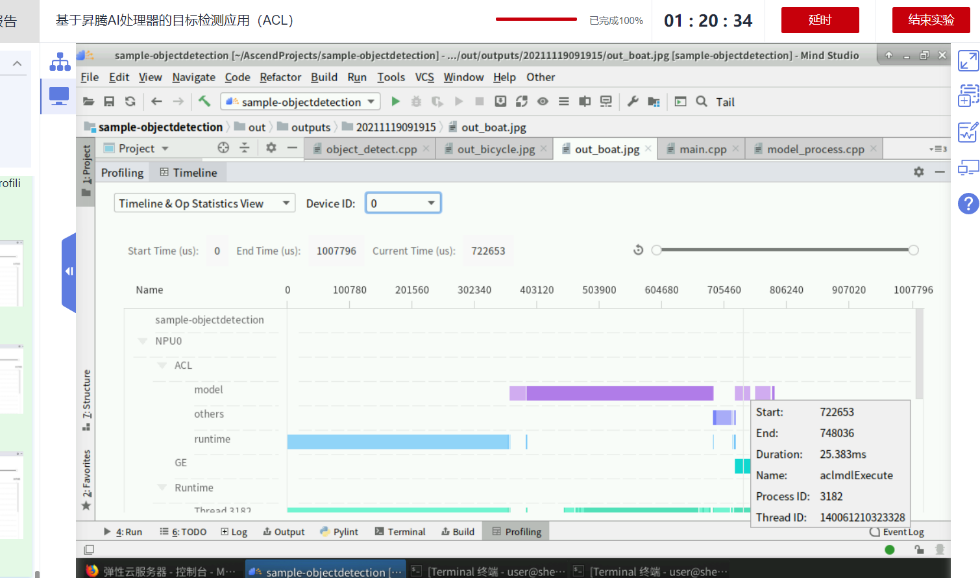


图5 第一张图片25.383ms

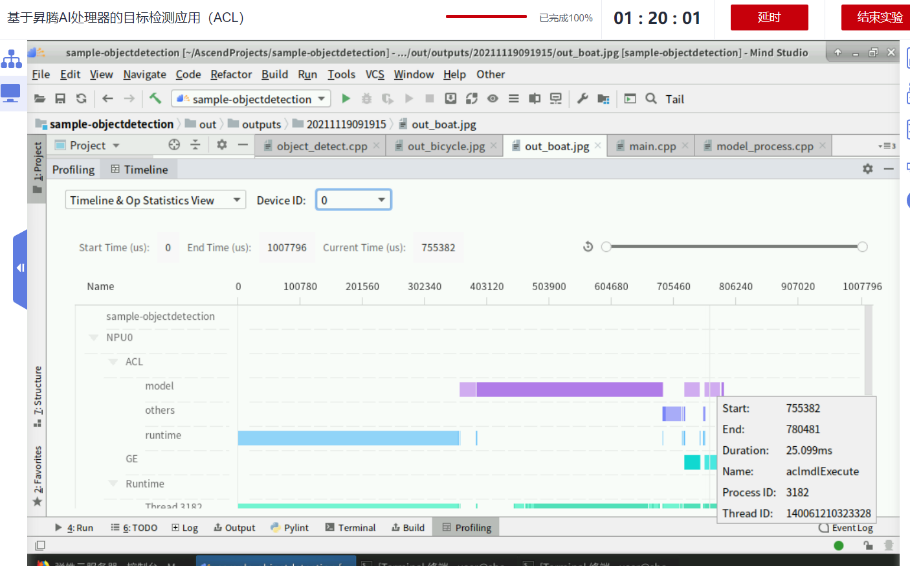


图6 第二张图25.099ms

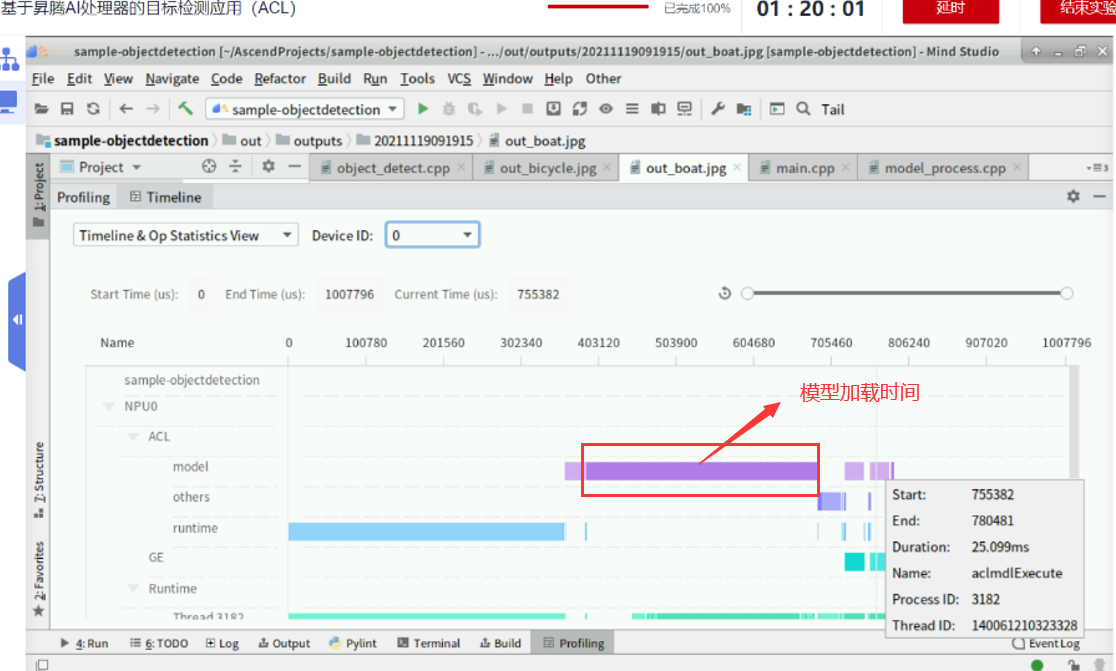


图7 模型从文件中加载

模型从文件加载进内存耗时：325.824ms

第一张图像推理耗时：25.383ms

第二张图像推理耗时：25.099ms

## 五、代码分析（10分）

（由于实验环境有时间限制，可以将代码下载到本地，参考华为实验手册与课程PPT分析）

源码工程：

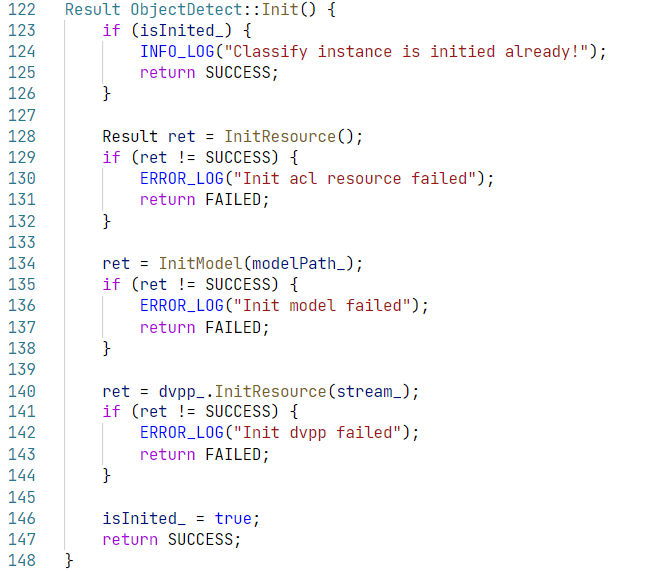
https://sandbox-experiment-resource-north-4.obs.cn-north-4.myhuaweicloud.com/shengteng-image-target/sample-objectdetection.zip

简单分析：

https://kqybjxvfo3.feishu.cn/docs/doccnr1WWpezXYWeeUykUKUpi8g#F8Ufjc

分析main.cpp可知，代码主要有五部分：初始化、数据预处理、执行推理、数据后处理、资源销毁。

1. 初始化（包括ASCENDCL资源初始化、模型初始化、dvpp资源初始化）（2分）



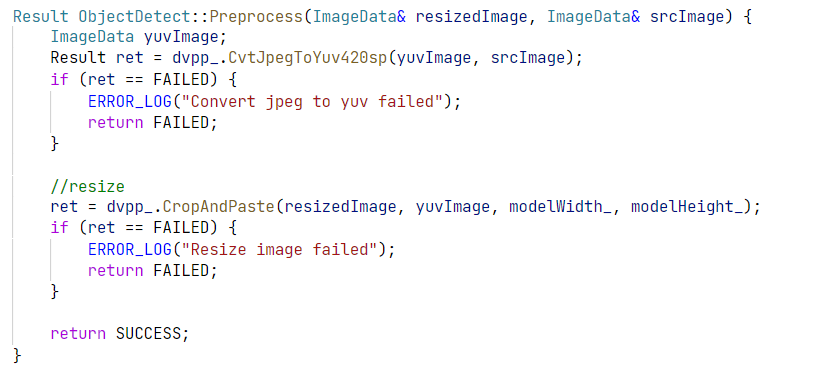
请查看InitResource函数的实现代码，ASCENDCL资源初始化时进行了哪些操作？

答：申请运行管理资源， setDevice、创建Context和stream、设置runMode。申请这些资源，是使用硬件算力的必备条件，可以理解为AscendCL开发必备的一个框架。

请查看InitModel函数的实现代码，模型初始化时进行了哪些操作？

答：将模型的参数初始化，初始化所有变量等

1. 数据预处理（2分）

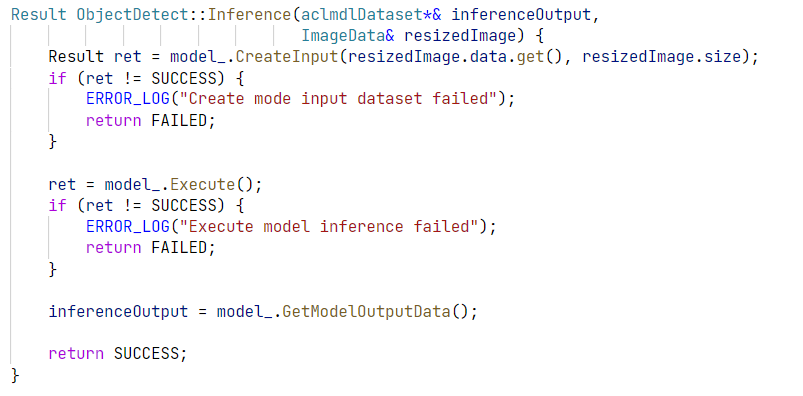


这一部分主要有两个封装好的函数，其作用分别是：

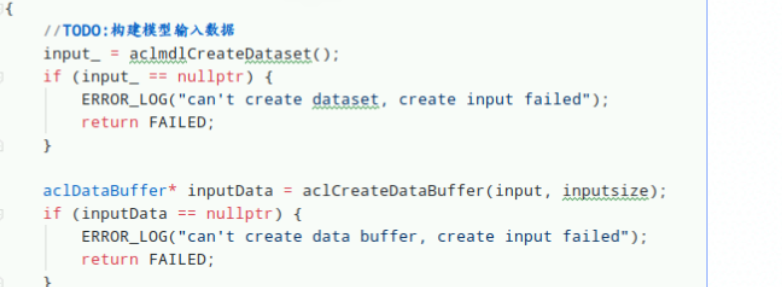
CvtJpegToYuv420sp：解码，dvpp把jpeg格式转化为Yuv420sp格式

CropAndPaste：缩放，dvpp把图片缩放为模型需要的长宽

1. 执行推理（2分）

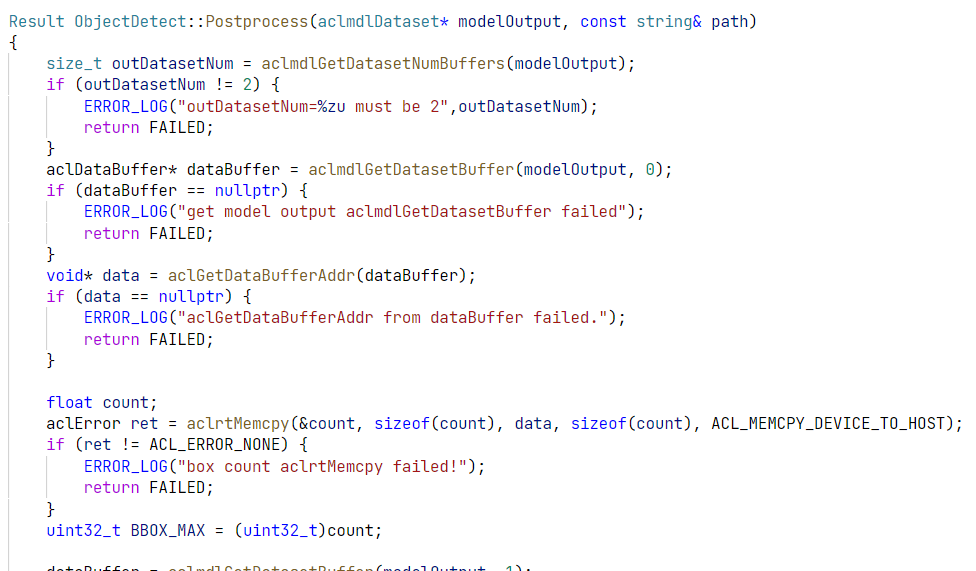


请查看CreateInput函数的实现代码，模型输入数据是怎样构造的？



答：获取aclmdlDataset类型的数据，并根据模型的描述信息获取模型的输出有几路数据

1. 数据后处理（2分）





上图为后处理的代码，它将模型输出的每一个检测结果（包括边界框点的位置、检测标签、置信度）在输入图像进行标注，简单分析一个检测结果占用多少字节？

答：标注了以下内容：

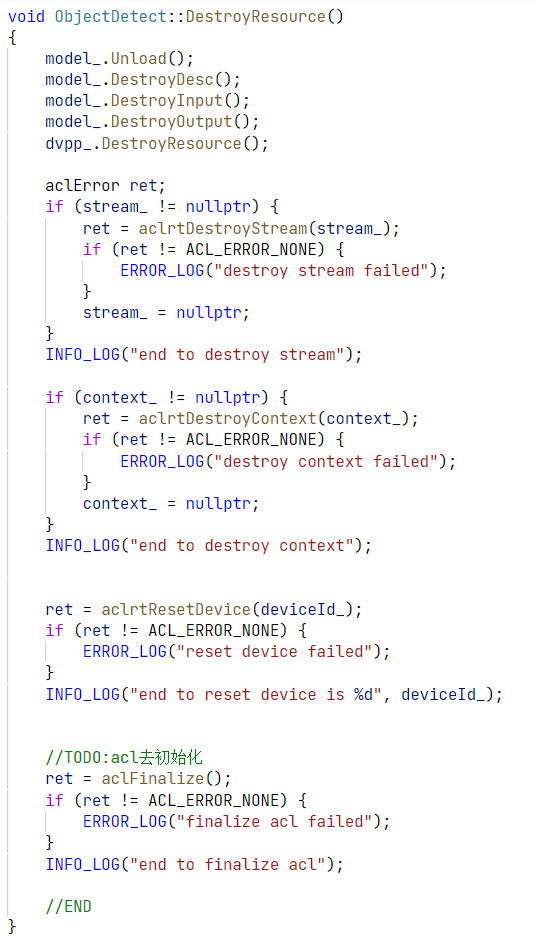
目标框的左上角坐标值x,y：int\*2=8Byte

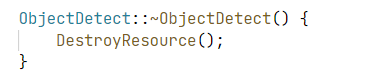
目标检测框的宽度和高度：int\*2=8Byte

以及检测出的物体类别索引：uint32\_t 4Byte

一共20 Byte

1. 资源销毁（2分）





~ObjectDetect()为析构函数，在该对象生命周期结束时自动调用，释放内存。其会调用DestroyResource()函数，该函数进行了哪些操作？

答：Device释放、Context释放、Stream释放、内存管理acl去初始化

## 六、实验感想

本次实验经历了这几个步骤：

1. 运行管理资源申请：用于初始化系统内部资源，固定的调用流程。
2. 加载模型文件并构建模型输出内存：加载模型文件到内存中，并根据内存中加载的模型获取模型的基本信息包含模型输入、输出数据的数据buffer大小，据此基本信息构建模型输出内存，为接下来的模型推理做好准备。
3. 获取本地图像数据并进行预处理：

于对原始模型的理解，在图像预处理时DVPP和AIPP分工处理：

DVPP：

1.解码：JPEG图片先解码为YUV420SP（输出128\*16对齐）

2.图像缩放：cropandpast（300,304）

3.输出图像数据类型：Uint8

AIPP：

1. 色域转换：YUV->BGR

2. 减均值

3. 抠出Crop（0, 0，300,300）

4. 图像数据类型转换：Uint8->FP16

（达芬奇推理的要求，模型最后实际的输入是fp16）)

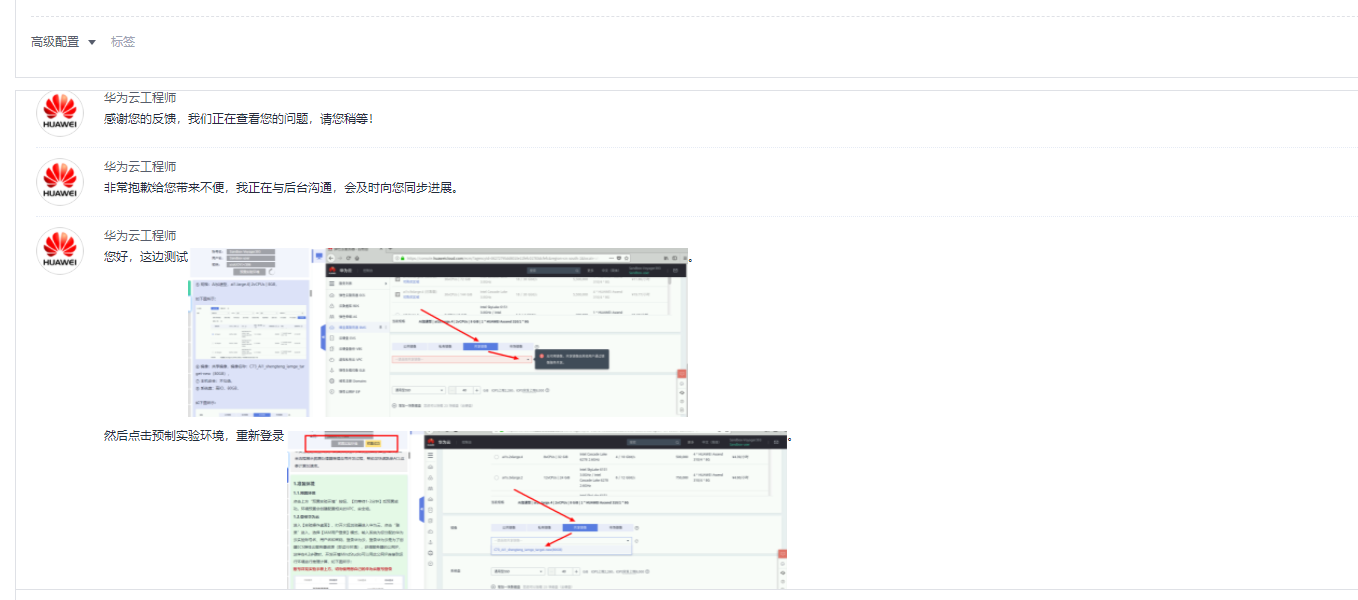
1. 构建模型输入数据：将经过预处理后的图像数据通过ACL接口，构建模型的输入数据，详细的调用流程在代码编写环节进行详细说明。
2. 进行模型推理：根据构建好的模型输入数据进行模型推理。
3. 解析模型推理结果：根据模型输出，解析目标检测的结果，得到图像数据中检测到的目标框，检测到的物体类别以及相似度。使用opencv将检测的结果标注在图像上并写入本地文件中。
4. 资源释放：程序运行结束时需要卸载模型，释放在设备侧申请的内存空间，对运行管理资源进行释放以及ACL去初始化。

## 附录：FAQ

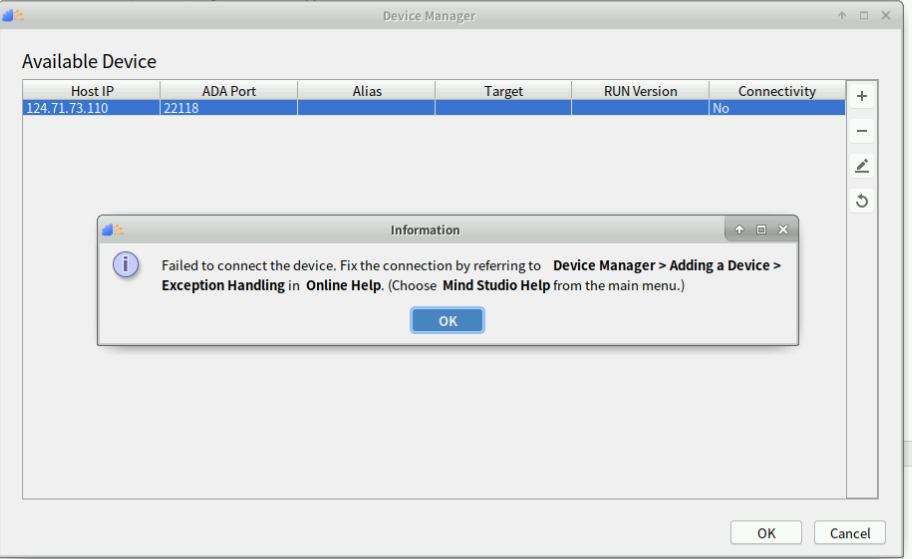
1. 找不到共享镜像？



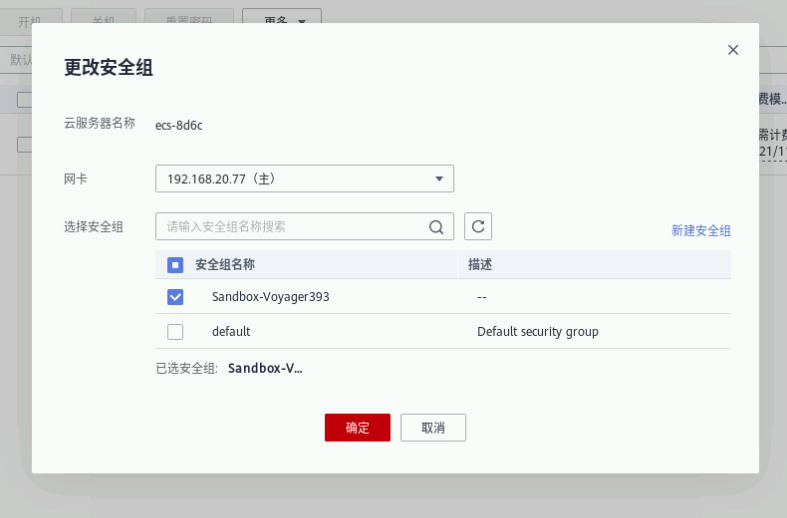
预置实验环境再重新登陆：

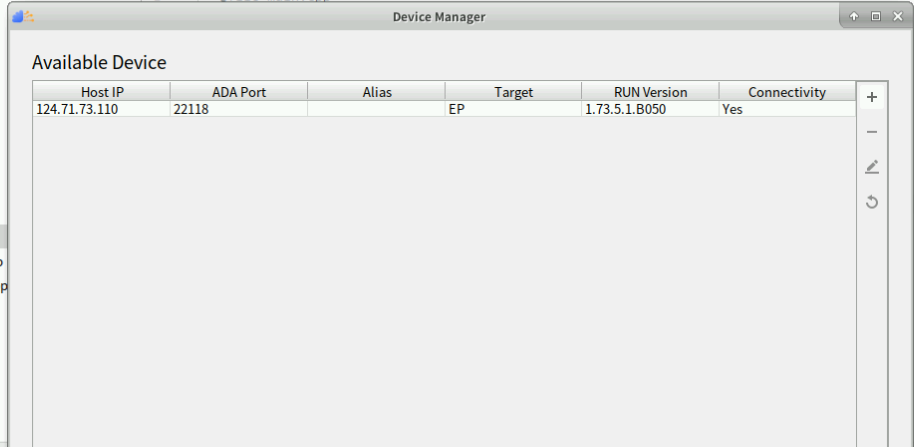


1. MindStudio无法连接云服务器？



更改安全组策略即可：





1. MindStudio运行项目，提示ERROR，并且错误信息中IP地址和端口与自己在Device Manager设置的不一致？

原因：MindStudio可能存在bug，由于某些原因，.idea/workspace.xml中hostip没有改变

解决一：将Device Manager中配置好的删掉，再重新配置一遍

解决二：按下图所示，选择AI Core，并Apply，再取消AI Core，再Apply

