

人工智能实验指导书

实验 2-深度学习实现花卉识别

2022 春

目录

1、实验目的.....	3
2、实验内容.....	3
3、实验原理.....	4
4、 本地基于不同框架实现花卉识别.....	5
4.1 基于 TensorFlow 框架的花卉识别	5
4.2 基于 MindSpore 框架的花卉识别	6
4.3 基于 PyTorch 框架的花卉识别	7
4.4 本地计算资源：高性能云服务中心	7
5、 云端 ModelArts 实现花卉识别.....	8
5.1 ModelArts 介绍.....	8
5.2 对象存储 OBS.....	9
5.3 ModelArts 实现花卉识别.....	9
5.3.1 使用预置算法实现.....	10
5.3.2 自定义算法实现.....	10
5.3.3 代金券的获取与使用.....	23
6、 实验结果提交.....	23

1、实验目的

- (1) 加深对神经网络相关知识的理解，如全连接神经网络、卷积神经网络等，熟悉其工作原理和实现。
- (2) 熟悉使用 TensorFlow、PyTorch、MindSpore 三种常用的深度学习框架，掌握深度学习开发环境和方法，了解不同框架的优缺点，能独立使用深度学习框架实现自定义的神经网络。
- (3) 了解产业趋势，掌握国产一站式 AI 开发平台 ModelArts 实现深度学习的一般流程，加深对国产深度学习框架的了解。

2、实验内容

与传统图像分类方法不同,卷积神经网络无需人工提取特征,可以根据输入图像,自动学习包含丰富语义信息的特征,得到更为全面的图像特征描述,可以很好地表达图像不同类别的信息。

本实验以小组形式完成（组队同实验一）。要求使用课程介绍的神经网络的知识，基于给定的数据集实现花卉识别。具体内容如下：

(1) 在本地分别使用 TensorFlow、MindSpore、PyTorch 训练深度学习模型实现花卉识别，包括但不限于课程介绍的各种模型。CPU 或 GPU 平台不做限制，操作系统不做限制，框架版本不做限制，开发语言不做限制（主流是 Python）；

(2) 在完成（1）的基础上，将本地的实现移植到云端 ModelArts 平台，即在 ModelArts 平台分别使用 TensorFlow、MindSpore、PyTorch 实现模型的训练，训练后的模型可下载到本地进行推理测试或者部署为线上服务实现在线预测图片。

注意：不能使用 model zoo 等定义好的模型或者其他预训练好的模型，要求必须自己一层层实现模型的定义。可以自行探索预训练模型 fine-tune 的效果，在实验报告或者答辩中可以将自定义的模型与其进行对比。模型在测试集上的精度非唯一评分标准，更注重模型设计、理解、实现。

解压实验包，在“花卉图像识别”目录可以看到如下子目录：

- flower_photos
- mindspore
- pytorch
- tensorflow
- TestImages
- readMe.md

flower_photos 是训练集数据，有 6 种图片，每种图片的数量不等。

TestImages 有 6 张测试图片，每个类别一张。

Tensorflow 文件夹中，lab2_flower_classify.ipynb 以记事本的形式，详细说明了数据处理、模型训练、推理预测等过程，需要补充完成模型定义等核心代码；lab2_flower_classify.py 是对应 python 代码，需要补充的代码以 Todo 的形式做了标注；lab2_flower_classify_tf_ModelArts 是 ModelArts 部署参考代码。

pytorch 和 mindspore 文件夹分别提供了在 ModelArts 进行在线部署的参考代码。

训练数据集包含 6 个类别的花卉图片，分别存在 6 个文件夹下，每个类别的图片数量和图片大小都不等。允许对训练集进行各种预处理，也可以另外找图片扩充训练集，但不可将测试集的 6 张图片用做训练，需要保证测试评估用的图片是没有参与过训练的。

- bee_balm
- blackberry_lily
- blanket_flower
- bougainvillea
- bromelia
- foxglove

类别	数量
Bee balm（蜂香薄荷）	66
Blackberry lily（黑莓百合花）	48
Blanket flower（天人菊）	49
Bougainvillea（叶子花）	128
Bromelia（凤梨花）	63
Foxglove（毛地黄）	162

对于有兴趣探索的同学，可以了解下开源数据集 oxford_flowers102（<https://www.robots.ox.ac.uk/~vgg/data/flowers/102/>）。

3、实验原理

该实验可以划分为数据处理、模型构建、模型评估与测试三个主要步骤。其中数据处理包括图像预处理、数据集划分两个部分；模型构建主要包括模型定义、模型训练以及模型部署三个部分；模型评估与测试主要包括读取花卉图像、运行模型推理进行图像特征提取，输出模型识别结果，评估模型质量。

因实验所用的开源环境官方会不定期更新，小版本众多，以最新的官方资料为准。同时因篇幅限制，指导书给出了参考资料，不再对步骤及配置做详细说明。

4、本地基于不同框架实现花卉识别

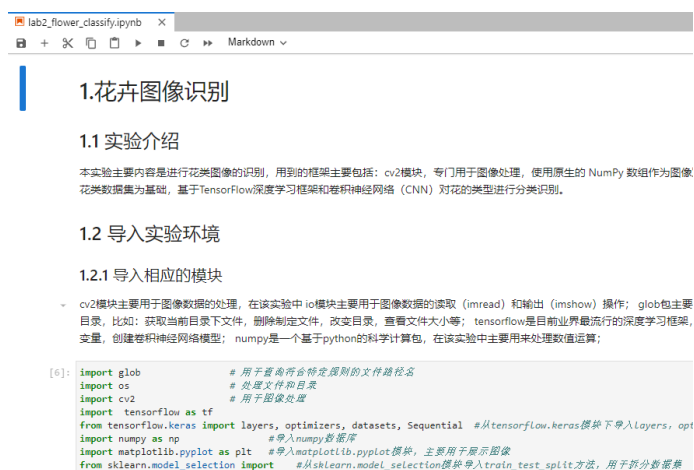
4.1 基于 TensorFlow 框架的花卉识别

TensorFlow 是谷歌于 2015 年开源的端到端深度学习框架，源自 Google Brain 内部孵化 DistBelief 项目，在工业界应用广泛。有 1.x 和 2.x 两个大版本，两个版本接口区别比较大，1.x 版本接口较混乱，建议使用 2.x 版本，具体安装、api 介绍、使用请参考 TensorFlow 官方教程：<https://tensorflow.google.cn/>

为了将主要精力集中在模型设计、超参数调整等核心内容，**本实验提供基于 TensorFlow2.x 版本的框架代码**，只需在此基础上自行完成模型设计、模型实现等步骤，然后进行模型训练和测试，鼓励有兴趣的同学不使用参考代码从零实现。指导书只给出主要要求说明，详细代码说明请查看 lab2_flower_classify.ipynb 文件。lab2_flower_classify.ipynb 文件最好使用 jupyter notebook 打开。

jupyter 的安装方式参考 <https://jupyter.org/install>；也可以使用“高性能云计算服务中心”平台的“Jupyterlab 1.1”打开，详见 4.4 节；也可以使用 ModelArts->开发环境->Notebook 打开。

打开后如下图，可以直接运行其中的代码。lab2_flower_classify.py 是对应的 Python 代码文件。



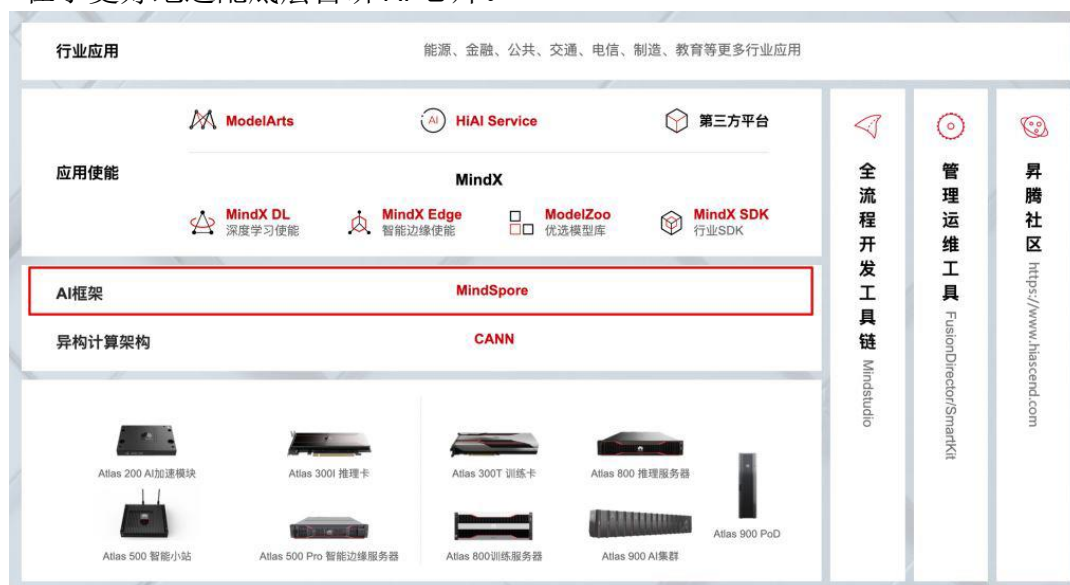
同学们定义模型，训练完成后，本地预测效果如下图所示：



4.2 基于 MindSpore 框架的花卉识别

MindSpore（昇思）是国内华为于 2019 年推出的最佳匹配昇腾 AI 处理器算力的全场景深度学习框架，2020 年正式开源，支持端、边缘、云全场景的计算框架，为数据科学家和算法工程师提供设计友好、运行高效的开发体验，推动人工智能软硬件应用生态繁荣发展。目前 MindSpore 支持在 EulerOS、Ubuntu、Windows 系统上安装，但是在 Windows 上只支持 CPU，Ubuntu 上同时支持 CPU 和 GPU、Ascend910，EulerOS 上只支持 Ascend910。

虽然市面上的 AI 框架众多，但 AI 计算框架发展还远未收敛，例如面向自动微分与张量计算的 AI 即时编译加速技术、面向超大规模神经网络的自动并行技术。华为昇腾 AI 全栈及 MindSpore 所处的位置如下图，MindSpore 框架的优势之一在于更好地适配底层自研 AI 芯片。



在 TensorFlow 实现花卉识别的基础上，按照 MindSpore 的使用方式，替换对应的 api，或者基于开源的 MindSpore 图像分类代码，进行修改实现花卉识别。具体使用和实现不做说明，可参考以下资料。

Mindspore 官网: <https://www.mindspore.cn/>

MindSpore 安装指南: <https://www.mindspore.cn/install>

MindSpore 教程: <https://www.mindspore.cn/tutorials/zh-CN/master/index.html>

MindSpore 实现花卉分类教程:

https://gitee.com/mindspore/course/blob/master/flowers_classification/flowers_classification.ipynb

4.3 基于 PyTorch 框架的花卉识别

PyTorch 由 Facebook 人工智能研究院于 2017 年开源的深度学习框架,对于有 Python 编程基础,对 Numpy 熟悉的人,上手非常快,社区资源多,在学术界广泛使用。

安装使用请参考: <https://pytorch.org/get-started/locally/>。

具体实现请阅读官方手册或开源代码完成

<https://pytorch.apachecn.org/#/docs/1.7/06>。

对于想深入学习的,推荐《动手学深度学习-pytorch 版本》
<https://zh.d2l.ai/index.html> 及配套视频。

4.4 本地计算资源:高性能云服务中心

为了更好的满足实验教学对计算资源的需要,实验与创新实践中心建设了“高性能计算云服务中心”。本地训练模型推荐使用“哈尔滨工业大学(深圳)高性能计算云服务中心”提供的 CPU 或 GPU 资源进行训练,平台地址为 <http://hpc.hitsz.edu.cn/>,使用统一认证登录,具体使用说明请参考 <http://hpc.hitsz.edu.cn/help/>。

因平台资源有限,无法像华为云等公有云平台提供无限资源,为了高效利用平台资源,避免部分同学占着资源不释放导致其他同学无法分配到资源,平台会对使用到的 CPU、GPU、存储等资源进行计费。每个账号有默认金额,足以满足实验要求,实验过程请合理使用集群资源,尽量少的使用虚拟机的方式,更多的使用 slurm 集群共享模式 (pytorch on slurm 教程 <https://wiki.hitsz.org/hpc-doc/pytorch/>),会使得计算资源的利用更高效。



5、云端 ModelArts 实现花卉识别

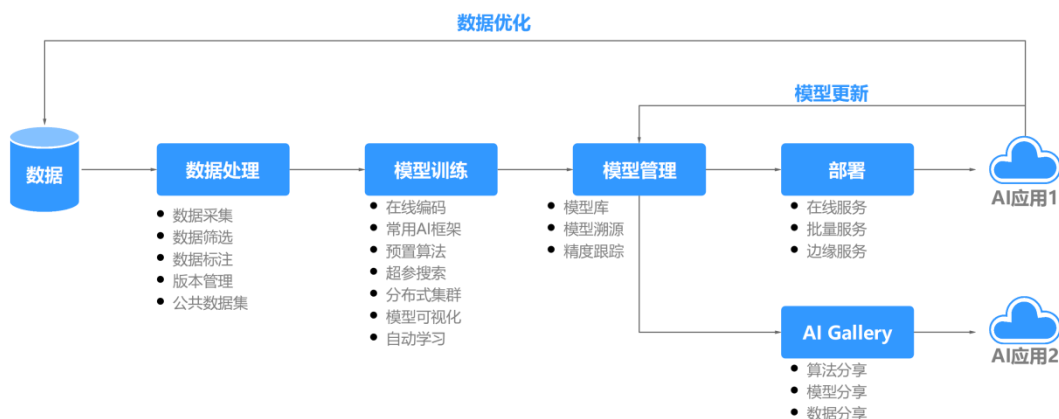
5.1 ModelArts 介绍

ModelArts 是云端面向 AI 开发者的一站式开发平台，提供海量数据预处理及半自动化标注、大规模分布式训练、自动化模型生成及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期 AI 工作流。



“一站式”是指 AI 开发的各个环节，包括数据处理、算法开发、模型训练、模型部署都可以在 ModelArts 上完成。从技术上看，ModelArts 底层支持各种异构计算资源，开发者可以根据需要灵活选择使用，而不需要关心底层的技术。同

时，ModelArts 支持 Tensorflow、PyTorch、MXNet、MindSpore 等主流开源的 AI 开发框架，也支持开发者使用自研的算法框架。



通过 ModelArts 技术结合花卉识别场景的实践，掌握 ModelArts 开发流程，了解 OBS 和 IAM 的基本应用；了解人工智能的应用，通过实践提升 ModelArts 开发的能力。使用 ModelArts 需要在华为云注册，参考《华为账号注册和华为云登录指南》进行账号注册与实名认证。

5.2 对象存储 OBS

不同于本地训练数据可以直接通过文件系统保存在磁盘，在 ModelArts 上实现模型训练，数据的存储方式一般使用对象存储。

华为的**对象存储服务**（Object Storage Service, OBS）是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力。对象存储是一类重要的存储方式，不同的云厂家有不同的产品，比如亚马逊的 S3、阿里云的 OSS。无论是哪家的产品，它们提供的服务是一样的，区别在于收费、性能、使用方式等。

OBS 系统和单个桶都没有总数据容量和对象/文件数量的限制，为用户提供了超大存储容量的能力，适合存放任意类型的文件。OBS 是一项面向 Internet 访问的服务，提供了基于 HTTP/HTTPS 协议的 Web 服务接口。这也限制了其访问方式不能像访问本地文件一样便捷。**OBS 不支持编辑，如果要修改文件只能重新上传替换之前的文件。**

ModelArts 需要用到 OBS 存储数据集、保存模型，在进行模型训练前先了解下，也可以在 ModelArts 实践过程中再做了解。

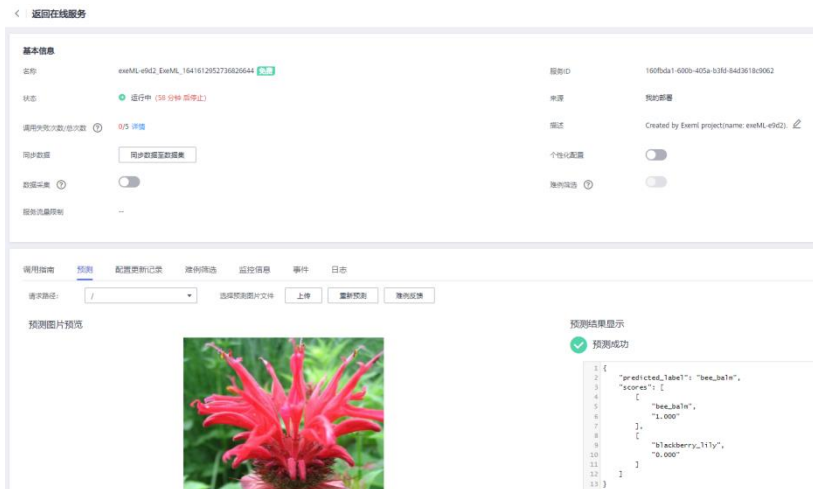
OBS 官方教程：<https://support.huaweicloud.com/obs/index.html>

5.3 ModelArts 实现花卉识别

ModelArts 官方教程：<https://support.huaweicloud.com/ModelArts/index.html>

ModelArts 支持使用预置算法、自定义算法和常用框架训练深度学习模型，

不同模式适用于不同开发对象和场景。我们需要在本地实现花卉识别的基础上，对代码进行适当改造，移植到 ModelArts 平台利用云端资源进行训练，并且可以直接部署成在线服务。在线服务效果如下图所示，能够上传图片并成功输出结果：



5.3.1 使用预置算法实现

ModelArts 内置了很多现成算法，只需上传数据集，给数据打上标签，就能自动训练、一键部署，大大降低了人工智能的门槛。鼓励大家体验，本实验不做要求，具体过程参考以下教程或自行在官网查找资料。

使用 AI Gallery 的订阅算法实现花卉识别

https://support.huaweicloud.com/bestpractice-ModelArts/ModelArts_10_0025.html

5.3.2 自定义算法实现

跟本地用 TensorFlow 等框架实现模型训练过程一致，需要自己写训练代码、推理代码，关联 OBS 内存的数据集进行训练。具体过程参考以下教程或自行在官网查找，指导书只对部分配置做了说明。

使用自定义算法构建模型（手写数字识别）

https://support.huaweicloud.com/bestpractice-modelarts/modelarts_10_0080.html

自定义训练算法的开发：

https://support.huaweicloud.com/engineers-modelarts/modelarts_23_0240.html

1. TensorFlow 框架的自定义训练和在线部署参考

以下以 TensorFlow 为例做简单说明。

(1) 修改代码以便迁移到云平台

可以参考 [tensorflow](#) 目录下的 `lab2_flower_classify_tf_modelarts.py` 文件。训练代码在本地 TensorFlow 训练的代码基础上进行两处修改：

- 1) 训练集和模型保存路径按以下方式获取（注意这里有 **train_url** 和 **data_url**，下面配置算法的时候会用到）：

```
import argparse
# 创建解析
parser = argparse.ArgumentParser(description="train flower classify",
                                formatter_class=argparse.ArgumentDefaultsHelpFormatter)
# 添加参数
parser.add_argument('--train_url', type=str,
                    help='the path model saved')
parser.add_argument('--data_url', type=str, help='the training data')
# 解析参数
args, unknown = parser.parse_known_args()

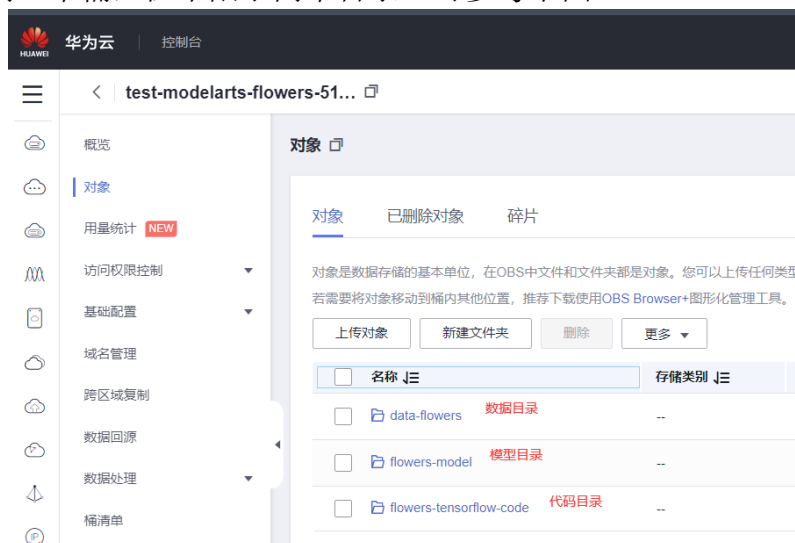
path = args.data_url
model_path = args.train_url
print(path)
print(model_path)
# /home/ma-user/modelarts/inputs/data_url_0/
# /home/ma-user/modelarts/outputs/train_url_0/
```

- 2) 使用 `model.save` 接口保存模型：

```
#输出模型的结构和参数量
model.summary()
model.save(model_path) #保存模型
```

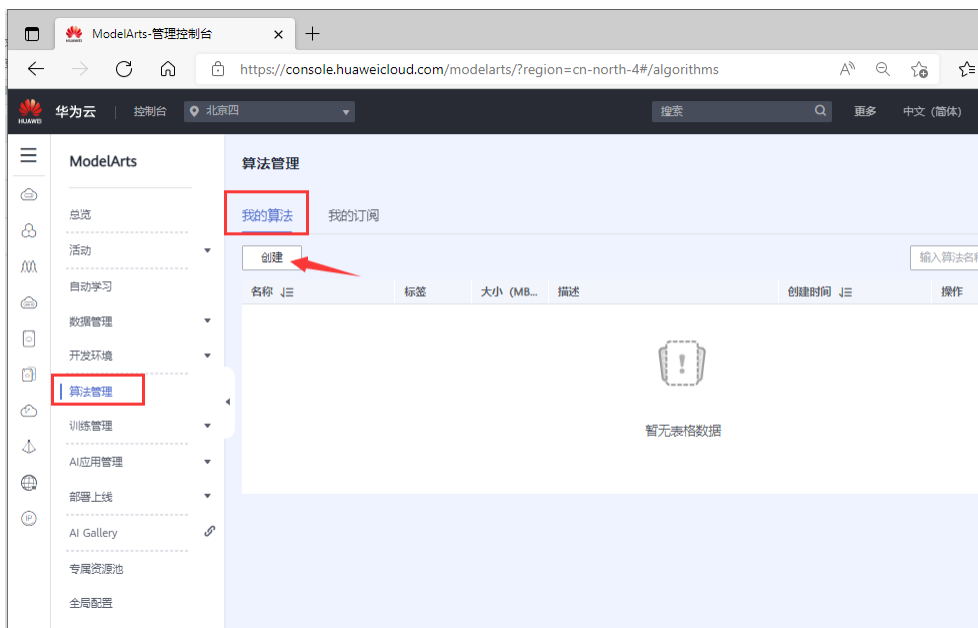
（2） 代码和数据上传 OBS

修改完代码后上传训练代码和数据集到 OBS 存储，训练代码和数据集需要放在同一个桶，但不限于同个目录。可参考下图：



（3） 创建算法

- 1) 在 ModelArts 管理控制台，进入“算法管理>我的算法”页面，单击左上角的“创建”：



2) 在创建算法页面，请参考下图填写相关信息：

创建算法 [返回我的算法](#)

名称

描述

创建方式 ☒ 自定义脚本 ☐ 自定义镜像

选择常用引擎创建训练作业。每个算法的代码目录最多支持1000个文件，文件深度不超过32，总大小不超过5GB。

引擎 ☐ 显示旧版引擎 [引擎升级说明](#)

代码目录 [选择](#) OBS上的代码存放路径

启动文件 [选择](#) 选择上传到OBS的代码

输入数据配置

为您的算法定义处理“输入数据”的参数，在您的算法代码中需要解析该参数获取到训练的数据集

映射名称 代码路径参数 ☒ 是 ☐ 否

[添加输入数据配置](#)

输出数据配置

为您的算法定义处理“输出数据”的参数，在您的算法代码中需要解析该参数获取到训练的输出生路径

映射名称 代码路径参数 ☒ 是 ☐ 否

[添加输出数据配置](#)

创建算法参考配置如下，添加 data_url 和 train_url 这两个配置，这两个变量对应的具体路径在创建训练任务时根据数据和模型在 OBS 存储的目录指定。

3) 确认无误后单击“提交”，算法创建成功：

算法管理

我的算法 我的订阅

创建

输入算法名

名称	标签	大小 (MB)	描述	创建时间	操作
algorithm-538d		0.01	--	2022/04/11 15:42:15 GMT+08:00	复制 删除 创建训练作业

(4) 训练模型

- 1) 在 ModelArts 管理控制台，进入“训练管理 > 训练作业 (New)”页面，单击左上角的“创建训练作业”：



- 2) 在训练作业配置页面配置信息：

< 训练作业

* 名称: trainjob-e73e

描述: 0/256

* 算法

我的算法 我的订阅

创建 快速创建

输入算法名称

名称	AI引擎	标签	描述	创建时间
algorithm-538d	TensorFlow tensorflow_2.1.0-cuda_10.1...	--		2022/04/11 15:42:15 GMT+08:00

训练输入

data_url: /test-modelarts-flowers-511c/data-flowers/ 数据集 数据存储位置

数据来源1

增加训练输入

train_url: /test-modelarts-flowers-511c/flowers-model/ 数据存储位置

输出数据1

增加训练输出

超参

增加超参

环境变量

增加环境变量

故障自动重启

★ 资源池 公共资源池 专属资源池

★ 资源类型 CPU GPU Ascend

★ 规格 [限时免费] GPU: 1*NVIDIA-V100(32GB) | CPU: 8 核 64GB 7... 获取输入数据大小

请确保已选择的规格有足够的磁盘空间下载输入文件

1、单个租户免费规格作业有一定限制。
2、免费规格的作业会在1个小时后自动停止，请勿下发运行时长超过1个小时的作业。
3、训练管理ModelArts免费算力不包含对象存储服务（OBS）存储资源费用，对象存储服务（OBS）计费标准详见如下链接：[对象存储服务（OBS）计费详情](#)。

☒ 我已阅读并同意以上内容

★ 计算节点个数 - 1 +

永久保存日志 ☐

日志30天后会被清理，打开按钮后可保存至指定OBS路径，您也可以在作业详情页下载全部日志至本地。

配置费用: ¥0.00/小时

提交

3) 确认无误后点“提交”：

名称/ID	训练类型	状态	创建时间	算法	描述	操作
trainjob-e73e 68858fa6-296b-443e-b4b0-38dd06811a29	训练作业	运行中	2022/04/11 15:51:23 GMT+08:00	algorithm-538d	--	删除 重建 终止

可以看到状态是“运行中”。

如果训练状态变为“运行失败”也可以通过查看训练日志找到问题：

名称/ID	训练类型	状态	创建时间	算法	描述	操作
trainjob-e73e 68858fa6-296b-443e-b4b0-38dd06811a29	训练作业	运行失败	2022/04/11 15:51:23 GMT+08:00	algorithm-538d	--	删除 重建 终止

trainjob-e73e

作业ID: 68858fa6-296b-443e-b4b0-38dd06811a29

作业状态: 运行失败

创建时间: 2022/04/11 15:51:23 GMT+08:00

运行时间: 00:00:31

描述: --

算法名称: algorithm-538d

代码目录: /test-modelarts-flowers-511c/flowers-tensorflow-code/

返回文件: /test-modelarts-flowers-511c/flowers-tensorflow-code/lab2_flow_classify_tf_modelarts.py

计算节点个数: 1

规格: [限时免费] GPU: 1*NVIDIA-V100(32GB) | CPU: 8 核 64GB

训练输入

输入路径	训练参数名称	本地路径 (训练参数值)
/test-modelarts-flowers-511c...	data_url	/home/ma-user/modellar...

训练输出

输出路径	训练参数名称	本地路径 (训练参数值)
------	--------	--------------

系统日志 | 当前已选文0/05MB

1 失败可能原因：模型运行中shape不匹配 | 解决方法：通过notebook调试，可参见文档。

```
350 File "/home/ma-user/anaconda/lib/python3.7/site-packages/tensorflow_core/python/keras/engine/training_v2.py", line 128, in run_one_epoch
351 batch_outs = execution_function(iterator)
352 File "/home/ma-user/anaconda/lib/python3.7/site-packages/tensorflow_core/python/keras/engine/training_v2_utils.py", line 98, in execution_function
353 distributed_function(input_fn)
354 File "/home/ma-user/anaconda/lib/python3.7/site-packages/tensorflow_core/python/eager/def_function.py", line 568, in __call__
355 result = self._call(*args, **kwargs)
356 File "/home/ma-user/anaconda/lib/python3.7/site-packages/tensorflow_core/python/eager/def_function.py", line 638, in __call__
357 return self._concrete_stateful_fn._filtered_call(canon_args, canon_kwargs) # pylint: disable=protected-access
358 File "/home/ma-user/anaconda/lib/python3.7/site-packages/tensorflow_core/python/eager/function.py", line 1661, in __call__
359 self._capture_inputs
360 File "/home/ma-user/anaconda/lib/python3.7/site-packages/tensorflow_core/python/eager/function.py", line 1692, in __call__
361 ctx, args, cancellation_manager=cancellation_manager))
362 File "/home/ma-user/anaconda/lib/python3.7/site-packages/tensorflow_core/python/eager/function.py", line 1645, in call
363 (ctx.execute)
364 File "/home/ma-user/anaconda/lib/python3.7/site-packages/tensorflow_core/python/eager/execute.py", line 67, in quick_execute
365 six.raise_from(core._status_to_exception(e.code, message), None)
366 File "test_training.py", line 3, in raise_from
367 tensorflow.python.framework.errors_impl.InvalidArgumentError: Incompatible shapes: [200,1] vs. [200,10,10]
368 [[node metrics/accuracy/Equal (defined at /modelarts/user-dir/flowers-tensorflow-code/lab2_flow_classify_tf_modelarts.py:183) ]]
369 [Op::Inference_distributed_function_266]
370
371 Function call stacks:
372 distributed_function
373
374 [2022-04-11T15:52:15+08:00] [ModelArts Service Log] exiting...
375 [2022-04-11T15:52:15+08:00] [ModelArts Service Log] exits with 1
376 [2022-04-11T15:52:16+08:00] [ModelArts Service Log] [sidecar] training is completed
377 time="2022-04-11T15:52:16+08:00" level=warning msg="get item by 319 line with error: value of repex is empty" file="cli.go:112" Command=analyze
378 Component=training-toolkit Platform=ModelArts-Service
379 time="2022-04-11T15:52:16+08:00" level=warning msg="the \"log-preview-size\" parameter exceeds the limit and will be set to the default value 5242880"
380 file="cli.go:192" Command=analyze Component=training-toolkit Platform=ModelArts-Service
381 [2022-04-11T15:52:16+08:00] [ModelArts Service Log] [sidecar] stop toolkit_obs_upload_by_channels.pid = 41 by signal SIGTERM
```

遇到问题需要修改代码重新上传 OBS，再把训练作业“重建”一下。

等待一段时间，当训练作业状态变为“已完成”时，即完成了模型训练过程。


名称/ID	训练类型	状态	创建时间	算法	描述	操作
trainjob-e73e-copy-2292 ee65d8a1-d098-4727-9968-fd05e9c701db	训练作业	已完成	2022/04/11 16:04:19 GMT+08:00	algorithm-538d	--	删除 重建 终止

这期间我们可以进入训练作业查看日志：



(5) 查看训练结果

训练完之后会在配置的输出路径（train_url）上生成模型文件 saved_model.pb 和数据文件 variables 目录：



variables 对象下的具体文件



(6) 应用部署

应用部署在线服务较复杂，部署过程比较耗时，调试也非常不方便。如果部署不成功，只能修改代码添加 print 语句，打印关键信息，再重新上传修改后的

代码到 OBS，再重新建 AI 应用（虽然 OBS 存储的文件已经更新了，但原有的 AI 应用的镜像用的旧的数据，要重新构建镜像），部署在线服务，然后看日志判断可能出错的地方。考虑到以上困难，实验对在线部署不做要求，建议将 ModelArts 上训练好的模型下载到本地进行评估测试。

下面是部署在线服务的步骤，供参考：

从训练作业导入的推理代码部署需要遵从模型包规范：

https://support.huaweicloud.com/engineers-modelarts/modelarts_23_0091.html。
即推理代码必须跟训练代码在同一个目录（如下图），且推理代码文件必须命名为 `customize_service.py`，再根据请求和返回参数修改配置文件 `config.json`。

对象

已删除对象

碎片

对象是数据存储的基本单位，在OBS中文件和文件夹都是对象。您可以上传任何类型（文本、图片、视频等）的文件，并在桶中对这些文件进行管理。[了解更多](#)

若需要将对象移动到桶内其他位置，推荐下载使用[OBS Browser+](#)图形化管理工具。

上传对象

新建文件夹

删除

更多 ▾

<input type="checkbox"/> 名称 ▾	存储类别 ▾	大小 ▾	加密状态 ▾
← 返回上一级			
<input type="checkbox"/> variables	--	--	--
<input type="checkbox"/> config.json  上传	标准存储	4.17 KB	未加密
<input type="checkbox"/> customize_service.py 	标准存储	4.82 KB	未加密
<input type="checkbox"/> lab2_flower_classify_tf_modelarts.py 	标准存储	5.63 KB	未加密
<input type="checkbox"/> saved_model.pb	标准存储	241.31 KB	未加密

模型配置文件 `config.json` 编写说明

https://support.huaweicloud.com/engineers-modelarts/modelarts_23_0092.html

TensorFlow 2.1 自定义保存模型加载与推理

https://support.huaweicloud.com/engineers-modelarts/modelarts_23_0301.html

回到 ModelArts 中准备创建 AI 应用：


华为云 | 控制台 | 北京四

ModelArts
 总览
 活动
 自动学习
 数据管理
 开发环境
 算法管理
 训练管理
 AI应用管理
AI应用
 模型转换

AI应用
 我的AI应用 | 我的订阅 | 云服务订阅AI应用
 创建 | 查找AI应用

AI应用名称	最新版本	状态	部署类型	版本数量
model-f7c6	0.0.1	正常	在线服务	1
model-4cbd	0.0.1	正常	在线服务	1
model-b546	0.0.1	正常	在线服务	1

创建AI应用

* 名称: model-9c32
 * 版本: 0.0.1
 描述:

* 元模型来源:
 从训练中选择
从对象存储服务 (OBS) 中选择
从容器镜像中选择
从模板中选择

训练作业 **训练作业 New**
 导入ModelArts训练作业中训练完成的模型。请在下方选择需要导入的训练作业。

* 选择训练作业: trainjob-e73e-copy-2292
☐ 动态加载

* AI引擎: TensorFlow | python3.6 **确认版本**
 推理代码: https://test-modelarts-flowers-511c.obs.myhuaweiclouds.com/flowers-model/customize_service.py

运行时依赖:

安装方式	名称	版本	约束
pip	numpy	1.15.0	当前版本以上
pip	Pillow	--	--

AI应用说明: [添加AI应用说明](#)

免费

立即创建

当状态为“正常”，表示 AI 应用创建成功：

AI应用

我的AI应用 | 我的订阅 | 云服务订阅AI应用

创建 | 查找AI应用

全部类型 | 请输入名称查询

AI应用名称	最新版本	状态	部署类型	版本数量	请求模式	创建时间	描述	操作
model-9c32	0.0.1	正常	在线服务	1	同步请求	2022/04/11 16:20:47 GMT+08:00	--	创建新版本 删除
model-f7c6	0.0.1	正常	在线服务	1	同步请求	2022/03/09 21:08:52 GMT+08:00	--	创建新版本 删除
model-4cbd	0.0.1	正常	在线服务	1	同步请求	2022/03/09 20:56:33 GMT+08:00	--	创建新版本 删除
model-b546	0.0.1	正常	在线服务	1	同步请求	2022/03/09 17:20:46 GMT+08:00	--	创建新版本 删除

单击操作列“部署>在线服务”，将模型部署为在线服务：

AI应用

我的AI应用 我的订阅 云服务订阅AI应用

创建 查找AI应用 全部类型 请输入名称查询

AI应用名称	最新版本	状态	部署类型	版本数量	请求模式	创建时间	描述	操作
model-9c32	0.0.1	正常	在线服务	1	同步请求	2022/04/11 16:20:47 GMT+08:00	--	创建新版本 删除
请输入版本查询								
版本	状态	部署类型	AI应用大小	模型来源	创建时间	描述	操作	
0.0.1	正常	在线服务	6.30 MB	自定义算法	2022/04/11 16:20:47 GMT+08:00	--	部署 发布 删除	
model-47c6	0.0.1	正常	在线服务	1	同步请求	2022/03/09 21:08:52 GMT+08:00	--	在线服务 批量服务 创建新版本 删除

1 2 3

< 部署

* 名称 service-62e0

是否自动停止 ? ☒

开启该选项后，在线服务的运行时间将在您选择的时间点后，自动停止。同时服务计费停止

☒ 1小时后 ☐ 2小时后 ☐ 4小时后 ☐ 6小时后 ☐ 自定义

描述

0/100

* 资源池 公共资源池 专属资源池

* 选择AI应用及配置

AI应用来源 我的AI应用 我的订阅

选择AI应用及版本 model-9c32 (同步请求) 0.0.1 (正常) C 分流 (%) ? - 100 +

计算节点规格 CPU: 2 核 8GB 这个选项比较便宜：CPU标准规格，满足大多数AI应用的运行和预测 计算节点个数 ? - 1 +

环境变量 ? ☒ 增加环境变量 为确保您的数据安全，在环境变量中，请勿输入敏感信息，如明文密码。

服务流量限制 ? ☐

配置费用 ¥0.80/小时 优先扣减免费套餐用量，了解更多 下一步

计费的，所以测试了赶紧关闭服务

ModelArts

在线服务 ?

温馨提示：状态为 运行中、部署中的服务正在产生费用，不使用时，请及时停止。若您已设置自动停止，请注意服务的剩余运行时间。

部署 删除 授权管理

名称/ID	状态	调用失败次数/总次数	来源	创建时间	描述	操作
service-62e0 e145c4fc-3a45-4a72-8d91-fab3fe878f72	部署中 (已耗尽%)	0/0	我的部署	2022/04/11 16:31:20 GMT+08:00	--	修改 预测 更多
service-62d4 a76d0940-d43f-4a00-b510-cc3620805d36	停止	1/1	我的部署	2022/03/09 21:12:41 GMT+08:00	--	修改 预测 更多
service-195e 8d982957-b61a-4d81-8508-d9e3836353ce	停止	1/1	我的部署	2022/03/09 20:58:13 GMT+08:00	--	修改 预测 更多
service-67a3 73835195-1775-4777-a8fb-fdd50ad48d7c	免费 停止	0/1	我的部署	2022/03/09 17:28:27 GMT+08:00	--	修改 预测 更多

1

等待部署成功，状态变成“运行中”：

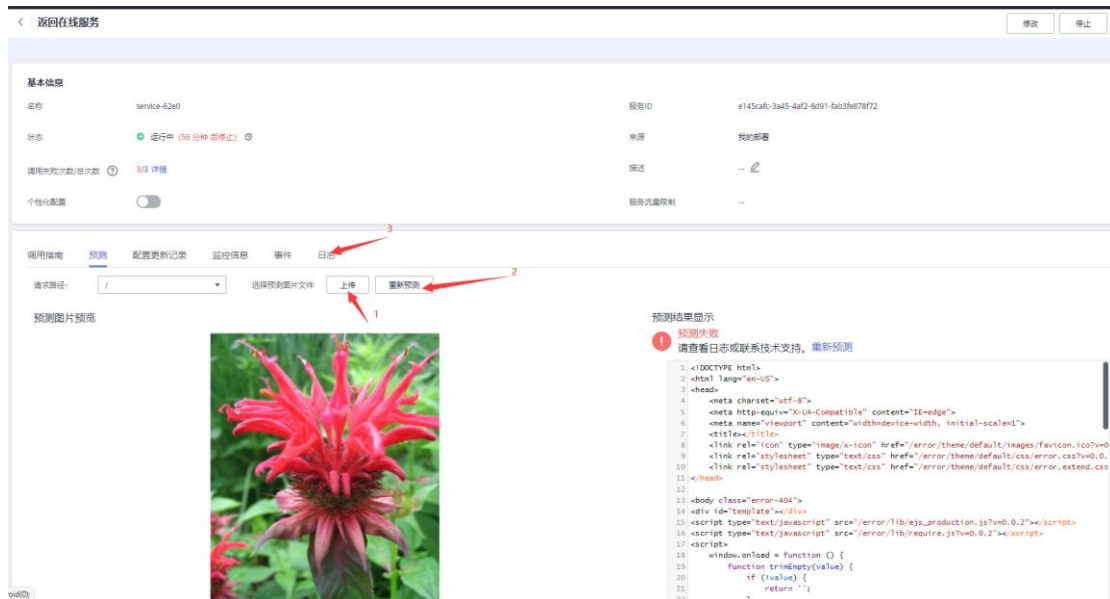
在线服务 ?

温馨提示：状态为 运行中、部署中的服务正在产生费用，不使用时，请及时停止。若您已设置自动停止，请注意服务的剩余运行时间。

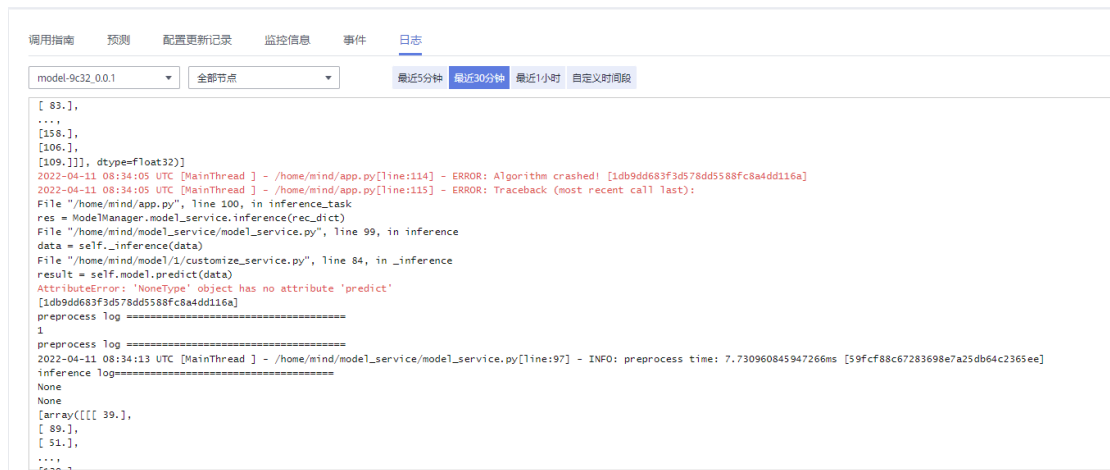
部署 删除 授权管理

名称/ID	状态	调用失败次数/总次数	来源	创建时间	描述	操作
service-62e0 e145c4fc-3a45-4a72-8d91-fab3fe878f72	运行中 (59 分钟后停止)	0/0	我的部署	2022/04/11 16:31:20 GMT+08:00	--	修改 预测 更多

点“预测”进入预测界面：



查看日志：



2. PyTorch 框架的自定义训练和在线部署参考

PyTorch 自定义脚本代码示例：

https://support.huaweicloud.com/engineers-modelarts/modelarts_23_0175.html。

训练代码在本地 TensorFlow 训练的代码基础上进行两处修改：

3) 训练集和模型保存路径按以下方式获取（注意这里有 **train_url** 和 **data_url**，下面配置算法的时候会用到）：

```
import argparse
# 创建解析
parser = argparse.ArgumentParser(description="train flower classify",
                                formatter_class=argparse.ArgumentDefaultsHelpFormatter)
# 添加参数
parser.add_argument('--train_url', type=str,
                    help='the path model saved')
parser.add_argument('--data_url', type=str, help='the training data')
# 解析参数
args, unknown = parser.parse_known_args()

path = args.data_url
model_path = args.train_url
print(path)
print(model_path)
# /home/ma-user/modelarts/inputs/data_url_0/
# /home/ma-user/modelarts/outputs/train_url_0/
```

4) 使用 model.save 接口保存模型:

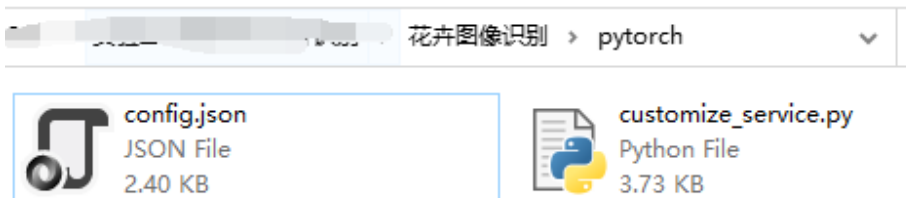
```
#输出模型的结构和参数量
model.summary()
model.save(model_path) #保存模型
```

对于 PyTorch 框架，如要在线部署，训练时必须采用 state_dict 的保存方式：

`torch.save(net.state_dict(), model_path + "/flower_mlp.pt")`

只保存模型权重参数，不保存模型结构，格式为 pt。

下发的资料中，Pytorch 目录下有 2 部署在线推理要用到的文件：



训练的时候如果用的 PyTorch1.8 的环境，在线部署时 AI 应用的推荐配置如下：

* 元模型来源 ?

从训练中选择
从模板中选择
从容器镜像中选择
从对象存储服务 (OBS) 中选择

此选项将导入您存储在对象存储服务 (OBS) 中的模型。目前仅支持从对象存储服务 (OBS) 导入 TensorFlow、MXNet、Caffe、PyTorch、Spark MLlib、Scikit Learn、XGBoost 模型。对于 Image 类型的模型建议您使用“从容器镜像中选择”的导入方式导入。您的模型文件需要存储于 model 目录下，选择模型时请选择 model 目录的上一级。如果您的模型代码，请放置 model 目录下，文件名需为 customize_service.py。模型包规范详情参见 模型包规范。

* 选择元模型 * AI引擎 PyTorch python3.7

☐ 动态加载 ?

推理代码 ?

运行时依赖

安装方式	名称	版本	约束	操作
pip	numpy	1.15.0	当前版本以上	删除
pip	Pillow			删除
pip	torch	1.8.0	当前版本以上	删除
pip	torchvision	0.8.1	当前版本以上	删除

增加

使用 Python3.7 的环境，通过 pip 的方式安装 PyTorch1.8，不然可能会因为版本差距过大在加载模型的时候出错。

运行环境 python3.7

AI引擎 PyTorch

模型来源 自定义算法

动态加载 否

AI应用说明 --

元模型来源 https://flower-img.obs.myhwclouds.com/pytorch

部署类型 在线服务

推理代码 https://flower-img.obs.myhwclouds.com/pytorch/model/customize_service.py

描述 --

推理环境

系统运行架构 X86 因训练用的版本是pytorch1.8.0，推理时所用版本需一致

推理加速卡类型 --

参数配置
运行时依赖
事件
使用约束

编辑
保存
?

安装方式	名称	版本	约束
pip	numpy	1.15.0	当前版本以上
pip	Pillow		--
pip	torch	1.8.0	当前版本以上
pip	torchvision	0.8.1	当前版本以上

3. MindSpore 框架的自定义训练和在线部署参考

MindSpore 在 ModelArts 上训练可以使用 CPU、GPU 或昇腾 910。实验推荐用 CPU 或 GPU，选择 AI 引擎：

MPI | mindspore_1.3.0-cuda_10.1-py_3.7-ubuntu_1804-x86_64。

创建训练作业提交如果出现以下错误：

algorithm-1c9c TensorFlow | tensorflow_2.1.0-cuda_10.1... tf-mnist
algorithm-af52 TensorFlow | tensorflow_2.1.0-cuda_10.1... tf-flower

data_url /flower-img/train/ 数据集 数据存储位置

训练数据

增加训练输入

train_url /flower-img/mindspore/model/ 数据存储位置

模型

增加训练输出

增加超参

增加环境变量

选择CPU或GPU都报错

公共资源池 专属资源池

CPU GPU Ascend

CPU: 2 核 8GB 50GB 获取输入数据大小

启动命令

选择 GPU，3200GB 存储的这个节点，如下图：

公共资源池 专属资源池

CPU GPU Ascend

GPU: 1*NVIDIA-V100(32GB) | CPU: 8 核 64GB 3200GB 获取输入数据大小

GPU: 8*NVIDIA-V100(32GB) | CPU: 72 核 512GB 3200GB

GPU: 1*NVIDIA-V100(32GB) | CPU: 8 核 64GB 3200GB

[限时免费] GPU: 1*NVIDIA-V100(32GB) | CPU: 8 核 64GB 780GB

由于 MindSpore 在 ModelArts 上部署比较复杂，建议将训练后生成的模型下载到本地，加载后进行预测。

MindSpore 在 ModelArts 部署只支持昇腾 310 芯片，而昇腾 310 支持的模型格式比较特殊，所以部署较为复杂，需要在训练的时候先生成 onnx 格式的模型，再通过模型转换功能转成 om 格式，再进行部署。这部分内容官方文档也没做详细说明。参考以下

ARM-Ascend 模板：

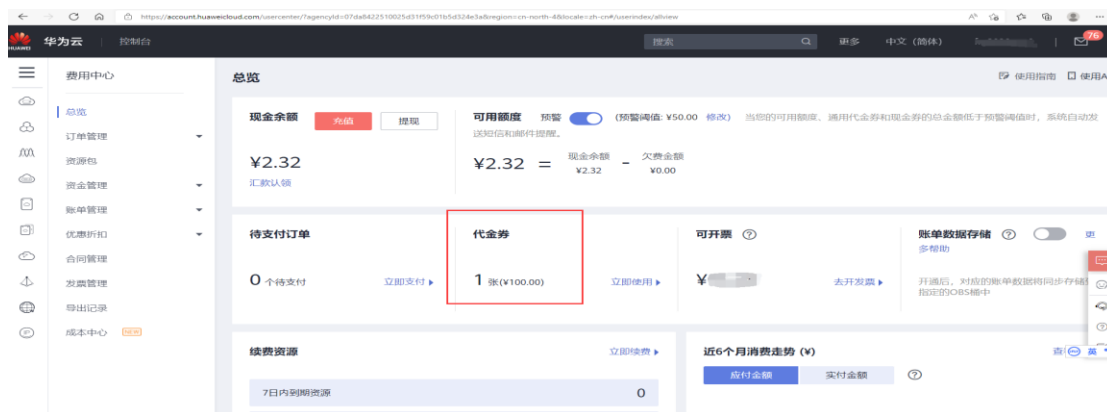
<https://support.huaweicloud.com/inference-modelarts/inference-modelarts-0072.html#inference-modelarts-0072>

图像分类 MindSpore 订阅算法：

<https://developer.huaweicloud.com/develop/aigallery/algorithm/detail?id=6b454013-dab9-4028-8c7a-47375067202c>

5.3.3 代金券的获取与使用

华为公司对本课程使用 ModelArts 资源提供了支持，将为每位同学发放华为云代金券。在实验开始前确保注册并实名认证华为云账号，助教将统一收集账号信息并申请代金券的发放，在实验课开始前及时查看账户是否有代金券，如无代金券及时上报给助教。



使用过程可在费用中心查看哪些服务进行了收费以及收费明细，如有则代金券可以抵扣费用。如对扣费有疑问可联系华为云客服。

账单详情 2022/01

应付金额 2022/01

¥0.53 = 现金支付 ¥0.00 + 代金券抵扣 ¥0.53 + 欠费金额 ¥0.00

汇总类别	总费用 (¥)	现金支付 (¥)	代金券抵扣 (¥)	欠费金额 (¥)
账单汇总	0.53	0.00	0.53	0.00
总计	0.53	0.00	0.53	0.00

展开趋势和汇总图表 更多维度按天按月分析，可到成本中心查看。

按产品汇总 2022/01

默认按照产品维度查看

套餐价 ¥0.56208212 = 优惠金额 ¥0.00000000 + 按零金额 ¥0.03208212 + 应付金额 ¥0.53

账单	账号	产品类型	产品	计费模式	账单类型	套餐价 (¥)	优惠金额 (¥)	按零金额 (¥)	应付金额 (¥)	现金支付 (¥)	代金券抵扣 (¥)	现金券抵扣 (¥)	储值卡抵扣 (¥)	欠费金额 (¥)
2022/01	hackenzheng_hit...	对象存储服务	云存储	按需	消费	0.01874881	0.00000000	0.00874881	0.01	0.00	0.01	0.00	0.00	0.00
2022/01	hackenzheng_hit...	ModelArts Mo...	modelarts虚拟...	按需	消费	0.54333331	0.00000000	0.02333331	0.52	0.00	0.52	0.00	0.00	0.00

注意：华为云代金券请用于实验用到的 ModelArts 和 OBS 等产品，代金券金额足以保证完成实验，每人限发一张，个人使用其他产品产生的扣费自行承担！！使用过程中务必注意系统算出来的计费金额，尽可能选择低计费的模式，比如 OBS 桶选择单 AZ，比如训练时能用单 GPU 卡完成的不要选择多 GPU 卡的节点。

计算资源一般会在任务结束后自动释放停止计费，但 OBS 里面的数据如果一直在会一直计费，在实验结束后请检查下存储，将模型等大文件进行删除避免一直计费。

6、实验结果提交

- (1) 每位同学书写自己完成功能的实验报告，小组合并后提交一份完整的实

验报告（注意标注每部分的作者名）；

（2） 代码以小组为单位提交最终完整版。每种框架的实现分目录保存，不需要提交生成的模型结构和模型参数文件；

（3） 提交截止时间见作业提交系统：<http://grader.tery.top:8000/#/courses>