

ModelArts

AI 工程师用户指南

文档版本 02

发布日期 2019-08-08

华为技术有限公司



版权所有 © 华为技术有限公司 2019。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI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。

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默示的声明或保证。

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。

目 录

1 AI 工程师如何使用 ModelArts.....	1
2 管理数据（推荐使用）	4
2.1 数据集简介.....	4
2.2 创建数据集.....	5
2.3 导入数据.....	7
2.3.1 导入操作.....	7
2.3.2 导入数据集的规范.....	9
2.3.2.1 OBS 目录.....	9
2.3.2.2 Manifest 文件.....	11
2.4 标注数据.....	19
2.4.1 图像分类.....	19
2.4.2 物体检测.....	25
2.4.3 文本分类.....	30
2.4.4 命名实体.....	33
2.4.5 声音分类.....	35
2.4.6 语音内容.....	38
2.4.7 语音分割.....	40
2.5 发布数据集.....	42
2.6 管理数据集版本.....	43
2.7 修改数据集.....	44
2.8 删 除数据集.....	45
3 数据管理.....	46
3.1 数据管理简介.....	46
3.2 数据标注.....	46
3.2.1 数据标注界面说明.....	46
3.2.2 创建数据标注作业.....	47
3.2.3 人工标注-图像分类.....	47
3.2.4 人工标注-物体检测.....	51
3.3 数据集.....	53
4 开发模型.....	57
4.1 Notebook.....	57
4.1.1 Notebook 简介.....	57

4.1.2 创建并打开 Notebook.....	58
4.1.3 访问 Notebook 并进行开发.....	61
4.1.3.1 使用 ModelArts 示例.....	61
4.1.3.2 使用 Convert to Python File 功能.....	63
4.1.3.3 使用 Sync OBS 功能.....	64
4.1.3.4 使用 Notebook 上传大文件.....	65
4.1.3.5 调用 ModelArts SDK.....	65
4.1.3.6 安装外部库和内核.....	66
4.1.4 启动或停止 Notebook.....	68
4.1.5 删除 Notebook.....	68
5 训练模型.....	69
5.1 模型训练简介.....	69
5.2 预置算法简介.....	70
5.3 创建训练作业.....	80
5.4 管理训练作业版本.....	86
5.5 查看作业详情.....	87
5.6 管理作业参数.....	91
5.7 管理 TensorBoard.....	92
6 管理模型.....	95
6.1 模型管理简介.....	95
6.2 (可选) 购买模型调优.....	95
6.3 导入模型.....	96
6.4 管理模型版本.....	99
6.5 模型二次调优.....	100
6.6 将模型发布至市场.....	100
6.7 模型模板.....	102
6.7.1 模板简介.....	102
6.7.2 模板模型包规范.....	105
6.7.3 输入输出模式说明.....	106
6.7.3.1 预置物体检测模式.....	107
6.7.3.2 预置图像处理模式.....	109
6.7.3.3 预置预测分析模式.....	109
6.7.3.4 未定义模式.....	112
6.8 转换模型.....	112
6.8.1 转换操作.....	112
6.8.2 模型输入目录规范.....	114
6.8.3 模型输出目录说明.....	115
6.8.4 转换模板.....	116
7 部署模型.....	118
7.1 模型部署简介.....	118
7.2 在线服务.....	118

7.2.1 部署为在线服务.....	118
7.2.2 查看服务详情.....	121
7.2.3 测试服务.....	122
7.2.4 访问在线服务.....	123
7.2.5 发布至市场.....	127
7.3 批量服务.....	128
7.3.1 部署为批量服务.....	128
7.3.2 查看批量服务预测结果.....	130
7.4 边缘服务.....	131
7.4.1 部署为边缘服务.....	131
7.4.2 访问边缘服务.....	132
7.5 修改服务.....	136
7.6 启动或停止服务.....	137
7.7 删除服务.....	138
8 AI 市场.....	139
9 资源池.....	142
10 权限管理.....	146
10.1 权限管理基本概念.....	146
10.2 创建并授权使用 ModelArts.....	151
10.3 创建 ModelArts 自定义策略.....	156
10.4 策略语法：细粒度策略.....	159
10.5 策略语法：RBAC.....	163
11 使用自定义镜像.....	166
11.1 自定义镜像简介.....	166
11.2 制作自定义镜像.....	167
11.3 上传镜像至 SWR.....	168
11.4 在 ModelArts 中使用自定义镜像.....	168
11.5 自定义镜像创建训练作业配置示例.....	171
11.6 附录：自定义镜像的补充说明.....	173
12 模型包规范.....	174
12.1 模型包规范介绍.....	174
12.2 模型配置文件编写说明.....	175
12.3 模型推理代码编写说明.....	181
13 将 DLS 数据迁移至 ModelArts.....	184
A 修订记录.....	188

1 AI 工程师如何使用 ModelArts

面向熟悉代码编写和调测，熟悉常见AI引擎的开发者，ModelArts不仅提供了在线代码开发环境，还提供了从数据准备、模型训练、模型管理到模型部署上线的端到端开发流程（即AI全流程开发），帮助您高效、快速的构建一个可用模型。

本文档介绍了如何在ModelArts管理控制台完成AI开发，如果您习惯使用API或者SDK进行开发，建议查看《[ModelArts SDK参考](#)》和《[ModelArts API参考](#)》获取帮助。

使用AI全流程开发的端到端示例，请参见[使用MXNet构建模型](#)和[使用Notebook构建模型](#)。

AI 全流程开发

ModelArts提供的AI全流程开发，兼容开发者的使用习惯，支持多种引擎和用户场景，使用自由度较高。下文介绍使用ModelArts平台，从数据准备到完成模型开发上线的全流程。针对开发者的其他场景，建议参考[ModelArts使用流程详解](#)。

图 1-1 AI 工程师的使用流程

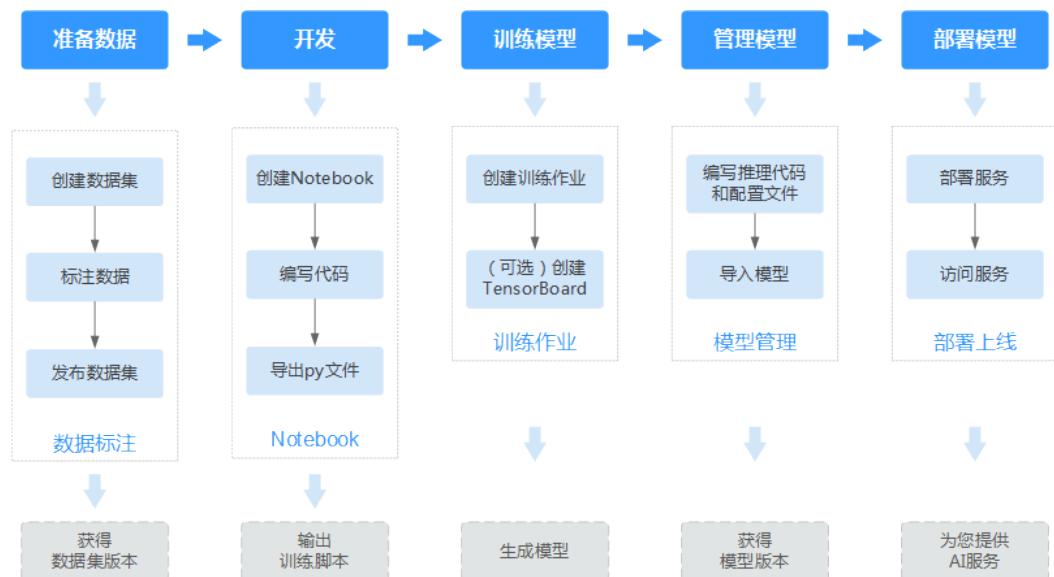


表 1-1 使用流程说明

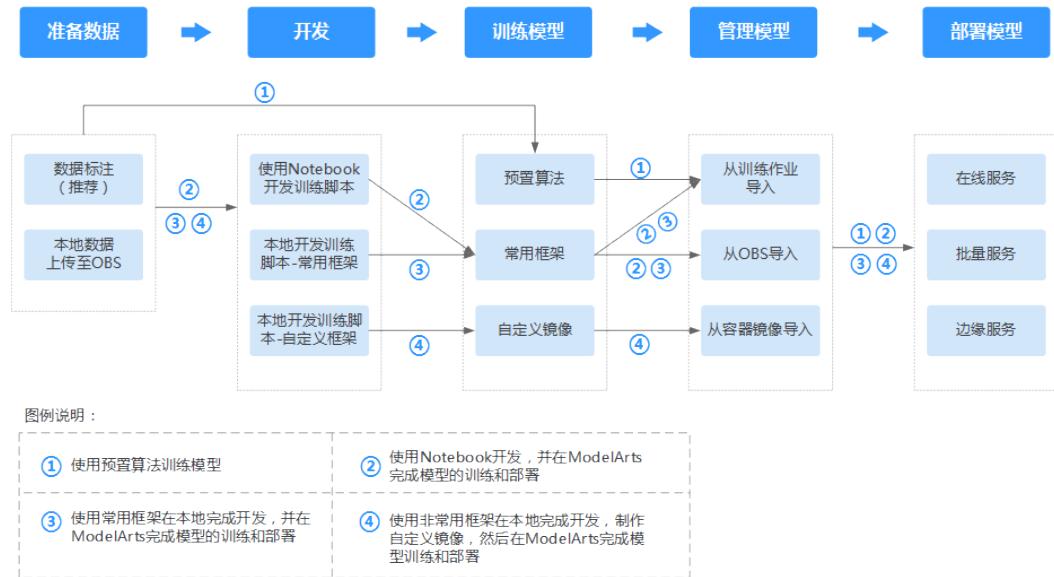
流程	子任务	说明	详细指导
准备数据	创建数据集	基于您的业务数据，您可以在ModelArts中创建数据集管理和预处理您的数据。	创建数据集
	标注数据	针对您创建的数据集，基于业务逻辑标注数据，对数据进行预处理，方便后续训练使用。数据标注的情况将影响模型训练效果。	标注数据
	发布数据集	数据标注完成后，将数据集发布。即可生成一个可以用于模型训练的数据集版本。	发布数据集
开发	创建Notebook	创建一个Notebook作为开发环境。	创建并打开Notebook
	编写代码	在已有的Notebook中编写代码，您也可以使用ModelArts示例，直接构建模型。	使用ModelArts示例
	导出“py”文件	编写完成的训练脚本，导出成“py”文件才可以用于后续的模型训练、模型管理等操作。	使用Convert to Python File功能
训练模型	创建训练作业	创建一个训练作业，选择可用的数据集版本，然后，上传并使用前面编写完成的训练脚本。训练完成后，将生成模型并存储至OBS中。	创建训练作业
	(可选) 创建TensorBoard	您可以通过创建TensorBoard查看模型训练过程，您可以通过TensorBoard提供信息，了解其模型情况，并对模型进行调整和优化。TensorBoard当前仅针对MXNet和TensorFlow引擎。	管理TensorBoard
管理模型	编写推理代码和配置文件	针对您生成的模型，建议您按照ModelArts提供的模型包规范，编写推理代码和配置文件，并将推理代码和配置文件存储至训练输出位置。	模型包规范介绍
	导入模型	将训练完成的模型导入至ModelArts，方便将模型部署上线。	导入模型
部署服务	部署服务	ModelArts支持将模型部署为在线服务、批量服务和边缘服务。	<ul style="list-style-type: none">● 部署为在线服务● 部署为批量服务● 部署为边缘服务

流程	子任务	说明	详细指导
	访问服务	服务部署完成后，针对在线服务和边缘服务，您可以访问并使用服务，针对批量服务，您可以查看其预测结果。	<ul style="list-style-type: none">● 访问在线服务● 访问批量服务● 访问边缘服务

ModelArts 使用流程详解

ModelArts平台提供了从数据准备到模型部署的AI全流程开发，针对每个环节，其使用时相对自由的。针对AI工程师，梳理了ModelArts使用流程详解，您可以选择其中一种方式完成AI开发。

图 1-2 使用流程详解



2 管理数据（推荐使用）

2.1 数据集简介

在ModelArts中，您可以在数据标注页面，完成数据导入、数据标注等操作，为模型构建做好数据准备。ModelArts以数据集为数据基础，进行模型开发或训练等操作。

标注类型

当前ModelArts支持如下7种类型的数据集。分别包含图片、音频和文本类别的。

- 图像分类：识别一张图片中是否包含某种物体。
- 物体检测：识别出图片中每个物体的位置及类别。
- 声音分类：对声音进行分类。
- 语音内容：对语音内容进行标注。
- 语音分割：对语音进行分段标注。
- 文本分类：对文本的内容按照标签进行分类处理。
- 命名实体：针对文本中的实体片段进行标注，如“时间”、“地点”等。

注意事项

- 目前ModelArts中创建的数据集，暂时无法应用于自动学习项目中。
- ModelArts存在“数据管理”和“数据标注”两块功能，都是用于管理数据集。由于“数据管理”模块将下线，推荐使用“数据标注”功能对数据集进行管理。如果您在“数据管理”模块存储了数据，请及时完成数据迁移。

数据集相关的操作

- **创建数据集**：创建一个新的数据集。
- **导入数据**：将本地或者OBS的数据导入数据集中。
- **标注数据**：针对7种不同类型的数据集，对数据进行标注。
- **发布数据集**：将标注后的数据集发布为新版本，以便应用于后续的模型构建。
- **管理数据集版本**：通过数据集版本查看演进过程。

- **修改数据集**: 修改数据集的基本信息。
- **删除数据集**: 删除数据集以释放资源。

2.2 创建数据集

在ModelArts进行数据管理时，首先您需要创建一个数据集，后续的操作，如标注数据、导入数据、数据集发布等，都是基于您创建和管理的数据集。

前提条件

- 数据标注功能需要获取访问OBS权限，在未进行委托授权之前，无法使用此功能。您可以在“数据标注”页面，单击“服务授权”，由具备授权的账号“同意授权”后，即可使用。
- 已创建用于存储数据的OBS桶及文件夹。
- 需要使用的数据已上传至OBS。

操作步骤

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 单击“创建数据集”，在创建数据集页面，参考**表2-1**填写信息，然后单击“创建”。

图 2-1 创建数据集

The screenshot shows the 'Create Dataset' interface. Key fields include:

- 名称:** dataset-demo
- 描述:** (empty, 0/256)
- 数据集输入位置:** /modelarts-demo1/dataset-input/
- 数据集输出位置:** /modelarts-demo1/dataset-output/
- 标注场景:** 物体 (selected)
- 标注类型:**
 - 图像分类:** Recognize if a picture contains a certain object. Options: cat (checked), dog.
 - 物体检测:** Detect the position and category of objects in a picture. Examples: Carl (Man), Women, Man, Women, Man.
- 添加标签集:**
 - 标签名称:** tag_demo_1
 - 请输入标签名称:** (empty)
 - 添加标签:** +

表 2-1 参数说明

参数名称	说明
名称	数据集的名称，名称只能是字母、数字、下划线、中划线或者中文字符组成的合法字符串。
描述	数据集的简要描述。
数据集输入位置	单击  选择数据集输入位置的OBS路径。
数据集输出位置	单击  选择数据集输出位置的OBS路径。 说明 “数据集输出位置”不能与“数据集输入位置”为同一路径。
标注场景	可选择“物体”、“音频”和“文本”三种标注场景。
标注类型	<ul style="list-style-type: none">● 标注场景为“物体”时<ul style="list-style-type: none">- 图像分类：识别一张图片中是否包含某种物体。- 物体检测：识别出图片中每个物体的位置及类别。● 标注场景为“音频”时<ul style="list-style-type: none">- 声音分类：对声音进行分类。- 语音内容：对语音内容进行标注。- 语音分割：对语音进行分段标注。● 标注场景为“文本”时<ul style="list-style-type: none">- 文本分类：对文本的内容按照标签进行分类处理。- 命名实体：针对文本中的实体片段进行标注，如“时间”、“地点”等。
添加标签集	<ul style="list-style-type: none">● 设置标签名称：在标签名称文本框中，输入标签名称。标签名称只能是字母、数字、下划线或中划线组成的合法字符串，不支持中文。长度为1~32字符。● 添加标签：单击  添加标签 添加标签。● 设置标签颜色：在每个标签右侧的标签颜色区域下，单击 ，然后在如下所示色板中选择颜色，或者直接输入十六进制颜色码进行设置。 

数据集创建完成后，系统自动跳转至数据集管理页面，针对创建好的数据集，您可以执行标注数据、发布、管理版本、修改、导入和删除等操作。

2.3 导入数据

2.3.1 导入操作

数据集创建完成后，一方面，可以直接从设置的数据集输入位置直接同步数据，另一方面，您还可以通过导入数据集的操作，导入更多数据。当前支持从OBS目录导入或从Manifest文件导入两种方式。

前提条件

- 已存在创建完成的数据集。
- 需导入的数据，已存储至OBS中。Manifest文件也需要存储至OBS。

导入方式

导入方式分为“OBS目录”和“Manifest文件”两种。

- OBS目录：指需要导入的数据集已提前存储至OBS目录中。此时需选择用户具备权限的OBS路径，且OBS路径内的目录结构需满足规范，详细规范请参见[OBS目录](#)。当前只有“图像分类”、“物体检测”、“文本分类”和“声音分类”4种类型的数据集，支持从OBS目录导入数据。其他类型只支持Manifest文件导入数据集的方式。
- Manifest文件：指数据集为Manifest文件格式，Manifest文件定义标注对象和标注内容的对应关系，且Manifest文件已上传至OBS中。Manifest文件压缩包的大小限制为最大8MB。Manifest文件的规范请参见[Manifest文件](#)。

从OBS目录导入

不同类型的数据集，导入操作界面的示意图存在区别，请参考界面信息了解当前类型数据集的示意图。当前操作指导以图像分类的数据集为例。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集所在行，单击操作列的“导入”。
3. 在“导入”对话框中，设置“导入方式”为“OBS目录”，然后在“OBS目录位置”中，设置数据存储的路径。然后单击“确定”。

图 2-2 导入数据集



导入成功后，数据将自动同步到数据集中。您可以在“数据标注”页面，单击数据集的名称，查看详细数据并进行数据标注。

从 Manifest 文件导入

不同类型的数据集，导入操作界面的示意图存在区别，请参考界面信息了解当前类型数据集的示意图。当前操作指导以物体检测类型的数据集为例。

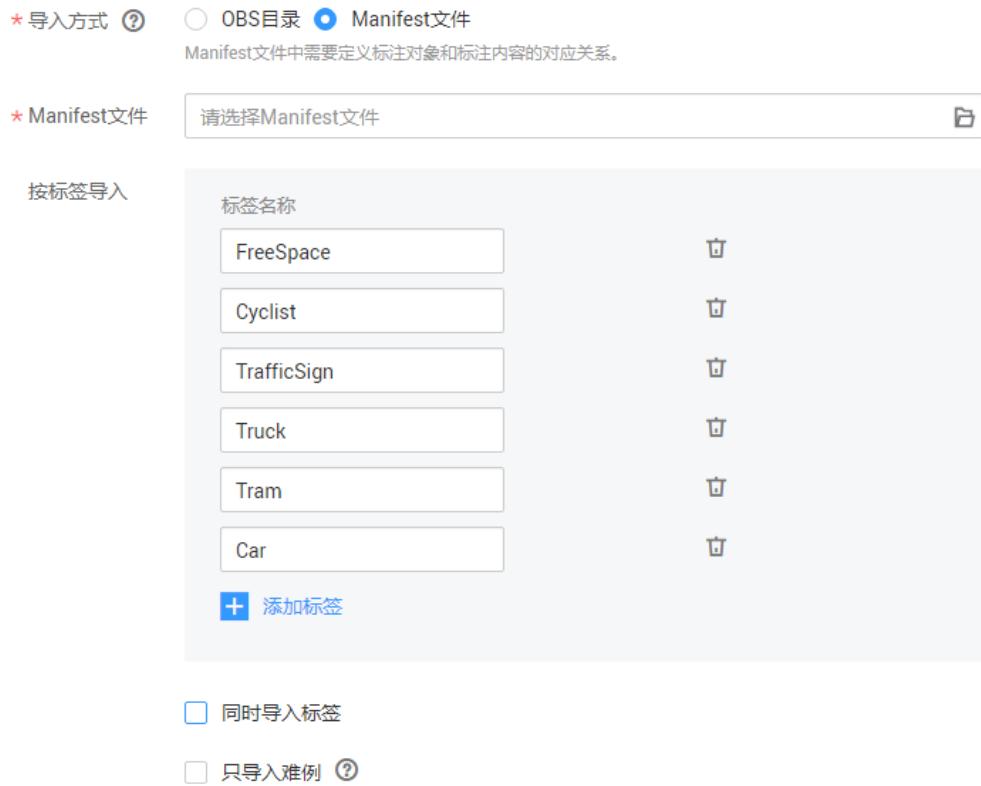
1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集所在行，单击操作列的“导入”。
3. 在“导入”对话框中，参考如下说明填写参数，然后单击“确定”。
 - “导入方式”：设置为“Manifest文件”。
 - “Manifest文件”：设置Manifest文件存储的OBS路径。

“按标签导入”：系统将自动获取此数据集的标签，您可以单击 添加，也可以单击标签右侧的 删除标签。此字段为可选字段，您也可以在导入数据集后，在标注数据操作时，添加或删除标签。

“同时导入标签”：勾选表示将Manifest文件中定义的标签一并导入ModelArts数据集中。

“只导入难例”：难例指Manifest文件中的“hard”属性，勾选此参数，表示此导入操作，只导入Manifest文件“hard”属性中数据信息。

图 2-3 导入数据集



导入成功后，数据将自动同步到数据集中。您可以在“数据标注”页面，单击数据集的名称，查看详细数据并进行数据标注。

2.3.2 导入数据集的规范

2.3.2.1 OBS 目录

导入数据集时，使用存储在OBS的数据时，数据的存储目录以及文件名称需满足 ModelArts 的规范要求。

当前只有“图像分类”、“物体检测”、“文本分类”和“声音分类”4种类型的数据集，支持从OBS目录导入数据。因此，如下内容，仅罗列此4种类型数据集规范。

图像分类

图像分类的数据要求将相同标签的图片放在一个目录里，并且目录名字即为标签名。

示例如下所示，其中Cat和Dog分别为标签名。

```
dataset-import-example
├── Cat
│   ├── 10.jpg
│   ├── 11.jpg
│   └── 12.jpg
└── Dog
    ├── 1.jpg
    ├── 2.jpg
    └── 3.jpg
```

- 如果导入位置为OBS， 用户需具备此OBS路径的读取权限。
- 只支持单标签。
- 只支持JPG、JPEG、PNG、BMP格式的图片。

物体检测

物体检测的简易模式要求用户将标注对象和标注文件存储在同一目录，并且一一对应，如标注对象文件名为“IMG_20180919_114745.jpg”，那么标注文件的文件名应为“IMG_20180919_114745.xml”。

物体检测的标注文件需要满足PASCAL VOC格式，格式详细说明请参见[表2-6](#)。

示例：

```
└──dataset-import-example
    ├── IMG_20180919_114732.jpg
    ├── IMG_20180919_114732.xml
    ├── IMG_20180919_114745.jpg
    ├── IMG_20180919_114745.xml
    ├── IMG_20180919_114945.jpg
    └── IMG_20180919_114945.xml
```

- 如果导入位置为OBS， 用户需具备此OBS路径的读取权限。
- 只支持JPG、JPEG、PNG、BMP格式的图片，且一次上传所有图片的总大小不能超过8MB。

文本分类

文本分类的标注对象和标注文件均为文本文件，并且以行数进行对应，如标注文件中的第一行表示的是标注对象文件中的第一行的标注。

例如，标注对象“COMMENTS_20180919_114745.txt”的内容如下所示。

```
手感很好，反应速度很快，不知道以后怎样
三个月前买了一个用的非常好果断把旧手机替换下来尤其在待机方面秒杀
没充一会电源怎么也会发热呢音量健不好用回弹不好
算是给自己的父亲节礼物吧物流很快下单不到24小时就到货了耳机更赞有些低音炮的感觉入耳很紧不会掉棒棒哒
```

标注文件“COMMENTS_20180919_114745_result.txt”的内容。

```
positive positive
negative
negative
positive
```

简易模式要求用户将标注对象和标注文件存储在同一目录，并且一一对应，如标注对象文件名为“COMMENTS_20180919_114745.txt”，那么标注文件名为“COMMENTS_20180919_114745_result.txt”。

数据文件存储示例：

```
└──dataset-import-example
    ├── COMMENTS_20180919_114732.txt
    ├── COMMENTS_20180919_114732_result.txt
    ├── COMMENTS_20180919_114745.txt
    ├── COMMENTS_20180919_114745_result.txt
    ├── COMMENTS_20180919_114945.txt
    └── COMMENTS_20180919_114945_result.txt
```

声音分类

声音分类的简易模式要求用户将相同标签的声音文件放在一个目录里，并且目录名字即为标签名。

示例：

```
dataset-import-example
├── Cat
│   ├── 10.wav
│   ├── 11.wav
│   └── 12.wav
└── Dog
    ├── 1.wav
    ├── 2.wav
    └── 3.wav
```

- 如果导入位置为OBS， 用户需具备此OBS路径的读取权限。

2.3.2.2 Manifest 文件

Manifest文件中定义了标注对象和标注内容的对应关系。此导入方式是指导入数据集时，使用Manifest文件。选择导入Manifest文件时，可以从本地文件系统导入或者从OBS导入。当从OBS导入Manifest文件时，需确保当前用户具备Manifest文件所在OBS路径的权限。

Manifest文件描述的是原始文件和标注信息，可用于标注、训练、推理场景。Manifest文件中也可以只有原始文件信息，没有标注信息，如用于推理场景，或用于生成未标注的数据集。Manifest文件需满足如下要求：

- Manifest文件使用UTF-8编码。文本分类的source数值可以包含中文，其他字段不建议使用中文。
- Manifest文件使用json lines格式（jsonlines.org），一行一个json对象。

```
{"source": "/path/to/image1.jpg", "annotation": ...}
{"source": "/path/to/image2.jpg", "annotation": ...}
{"source": "/path/to/image3.jpg", "annotation": ...}
```

为了说明方面，下面的Manifest例子格式化为多行的json对象。
- Manifest文件可以由用户、第三方工具或ModelArts数据标注生成，其文件名没有特殊要求，可以为任意合法文件名。为了ModelArts系统内部使用方便，ModelArts数据标注功能生成的文件名由如下字符串组成：“DatasetName-VersionName.manifest”。例如，“animal-v201901231130304123.manifest”。

图片分类

```
{
  "source": "s3://path/to/image1.jpg",
  "usage": "TRAIN",
  "id": "0162005993f8065ef47eefb59d1e4970",
  "annotation": [
    {
      "type": "modelarts/image_classification",
      "name": "cat",
      "property": {
        "color": "white",
        "kind": "Persian cat"
      },
      "hard": "true",
      "annotated-by": "human",
      "creation-time": "2019-01-23 11:30:30"
    }
  ]
}
```

```
    },
    {
        "type": "modelarts/image_classification",
        "name": "animal",
        "annotated-by": "modelarts/active-learning",
        "confidence": 0.8,
        "creation-time": "2019-01-23 11:30:30"
    }
],
"inference-loc": "/path/to/inference-output"
}
```

表 2-2 字段说明

字段	是否必选	说明
source	是	被标注对象的URI。数据来源的类型及示例请参考 表 2-3 。
usage	否	默认为空，取值范围： <ul style="list-style-type: none">● TRAIN：指明该对象用于训练。● EVAL：指明该对象用于评估。● TEST：指明该对象用于测试。● INFERENCE：指明该对象用于推理。 如果没有给出该字段，则使用者自行决定如何使用该对象。
annotation	否	如果不设置，则表示未标注对象。annotation值为一个对象列表，详细参数请参见 表2-4 。
inference-loc	否	当此文件由推理服务生成时会有该字段，表示推理输出的结果文件位置。

表 2-3 数据来源类型

类型	示例
OBS	"source": "s3://path-to-jpg"
Content	"source": "content://I love machine learning"

表 2-4 annotation 对象说明

字段	是否必选	说明
type	是	标签类型。取值范围为： <ul style="list-style-type: none">● image_classification：图像分类● text_classification：文本分类● text_entity：文本命名实体● object_detection：对象检测● audio_classification：声音分类● audio_content：声音内容● audio_segmentation：声音起止点
name	是/否	对于分类是必选字段，对于其他类型为可选字段，本示例为图片分类名称。
property	否	包含对标注的属性，例如本示例中猫有两个属性，颜色（color）和品种（kind）。
hard	否	表示是否是难例。“True”表示该标注是难例，“False”表示该标注不是难例。
annotated-by	否	默认为“human”，表示人工标注。
creation-time	否	创建该标注的时间。是用户写入标注的时间，不是Manifest生成时间。
confidence	否	表示机器标注的置信度。范围为0~1。

文本分类

```
{  
    "source": "content://I like this product ",  
    "id": "XGDVGS",  
    "annotation": [  
        {  
            "type": "modelarts/text_classification",  
            "name": "positive",  
            "annotated-by": "human",  
            "creation-time": "2019-01-23 11:30:30"  
        }  
    ]  
}  
{  
    "source": "content://I do not want to use it",  
    "annotation": [  
        {  
            "type": "modelarts/text_classification",  
            "name": "negative",  
            "annotated-by": "human",  
            "creation-time": "2019-01-23 11:30:30"  
        }  
    ]  
}
```

content字段是指被标注的文本（UTF-8编码，可以是中文），其他参数解释与[图片分类](#)相同，请参见[表2-2](#)。

文本命名实体

```
{  
    "source": "content://Michael Jordan is the most famous basketball player in the world.",  
    "usage": "TRAIN",  
    "annotation": [  
        {  
            "type": "modelarts/text_entity",  
            "name": "Person",  
            "property": {  
                "@modelarts:start_index": 0,  
                "@modelarts:end_index": 14  
            },  
            "annotated-by": "human",  
            "creation-time": "2019-01-23 11:30:30"  
        },  
        {  
            "type": "modelarts/text_entity",  
            "name": "Category",  
            "property": {  
                "@modelarts:start_index": 34,  
                "@modelarts:end_index": 44  
            },  
            "annotated-by": "human",  
            "creation-time": "2019-01-23 11:30:30"  
        }  
    ]  
}
```

“source”、“usage”、“annotation”等参数说明与[图片分类](#)一致，详细说明请参见[表2-2](#)。

其中，property的参数解释如[表2-5](#)所示。例如，当“"source": "content://Michael Jordan"”时，如果要提取“Michael”，则对应的“start_index”为“0”，“end_index”为“7”。

表 2-5 property 参数说明

参数名	数据类型	说明
@modelarts:start_index	Integer	文本的起始位置，值从0开始，包括start_index所指的字符。
@modelarts:end_index	Integer	文本的结束位置，但不包括end_index所指的字符。

物体检测

```
{  
    "source":  
    "s3://path/to/image1.jpg",  
    "usage":  
    "TRAIN",  
    "annotation": [  
        {  
            "type": "modelarts/object_detection",  
            "annotation-loc": "s3://path/to/annotation1.xml",  
            "annotation-format": "PASCAL VOC",  
            "annotated-by": "human",  
            "creation-time": "2019-01-23 11:30:30"  
        }  
    ]  
}
```

- “source”、“usage”、“annotation”等参数说明与[图片分类](#)一致，详细说明请参见[表2-2](#)。
- “annotation-loc”：对于物体检测是必选字段，对于其他类型是可选字段，标注文件的存储路径。
- “annotation-format”：描述标注文件的格式，可选字段，默认为“PASCAL VOC”。目前只支持“PASCAL VOC”。

表 2-6 PASCAL VOC 格式说明

字段	是否必选	说明
folder	是	表示数据源所在目录。
filename	是	被标注文件的文件名。
size	是	表示图像的像素信息。 <ul style="list-style-type: none">● width：必选字段，图片的宽度。● height：必选字段，图片的高度。● depth：必选字段，图片的通道数。
segmented	是	表示是否用于分割。
object	是	表示物体检测信息，多个物体标注会有多个object体。 <ul style="list-style-type: none">● name：必选字段，标注内容的类别。● pose：必选字段，标注内容的拍摄角度。● truncated：必选字段，标注内容是否被截断（0表示完整）。● occluded：必选字段，标注内容是否被遮挡（0表示未遮挡）。● difficult：必选字段，标注目标是否难以识别（0表示容易识别）。● confidence：可选字段，标注目标的置信度，取值范围0-1之间。● bndbox：必选字段，标注框的类型，可选值请参见表2-7。

表 2-7 标注框类型描述

type	形状	标注信息
point	点	点的坐标。 <x>100<x> <y>100<y>

type	形状	标注信息
line	线	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>200<y2>
bndbox	矩形框	左上和右下两个点坐标。 <xmin>100<xmin> <ymin>100<ymin> <xmax>200<xmax> <ymax>200<ymax>
polygon	多边形	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>100<y2> <x3>250<x3> <y3>150<y3> <x4>200<x4> <y4>200<y4> <x5>100<x5> <y5>200<y5> <x6>50<x6> <y6>150<y6>
circle	圆形	圆心坐标和半径。 <cx>100<cx> <cy>100<cy> <r>50<r>

示例：

```
<annotation>
  <folder>test_data</folder>
  <filename>260730932.jpg</filename>
  <size>
    <width>767</width>
    <height>959</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>point</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <occluded>0</occluded>
```

```
<difficult>0</difficult>
<point>
    <x1>456</x1>
    <y1>596</y1>
</point>
</object>
<object>
    <name>line</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <occluded>0</occluded>
    <difficult>0</difficult>
    <line>
        <x1>133</x1>
        <y1>651</y1>
        <x2>229</x2>
        <y2>561</y2>
    </line>
</object>
<object>
    <name>bag</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <occluded>0</occluded>
    <difficult>0</difficult>
    <bndbox>
        <xmin>108</xmin>
        <ymin>101</ymin>
        <xmax>251</xmax>
        <ymax>238</ymax>
    </bndbox>
</object>
<object>
    <name>boots</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <occluded>0</occluded>
    <difficult>0</difficult>
    <polygon>
        <x1>373</x1>
        <y1>264</y1>
        <x2>500</x2>
        <y2>198</y2>
        <x3>437</x3>
        <y3>76</y3>
        <x4>310</x4>
        <y4>142</y4>
    </polygon>
</object>
<object>
    <name>circle</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <occluded>0</occluded>
    <difficult>0</difficult>
    <circle>
        <cx>405</cx>
        <cy>170</cy>
        <r>100</r>
    </circle>
</object>
</annotation>
```

声音分类

```
{
"source": "s3://path/to/pets.wav",
"annotation": [
```

```
{  
    "type": "modelarts/audio_classification",  
    "name": "cat",  
    "annotated-by": "human",  
    "creation-time": "2019-01-23 11:30:30"  
}  
]  
}
```

“source”、“usage”、“annotation”等参数说明与[图片分类](#)一致，详细说明请参见[表2-2](#)。

语音内容

```
{  
    "source": "s3://path/to/audiol.wav",  
    "annotation": [  
        {  
            "type": "modelarts/audio_content",  
            "property": {  
                "@modelarts:content": "Today is a good day."  
            },  
            "annotated-by": "human",  
            "creation-time": "2019-01-23 11:30:30"  
        }  
    ]  
}
```

- “source”、“usage”、“annotation”等参数说明与[图片分类](#)一致，详细说明请参见[表2-2](#)。
- “property”中的“@modelarts:content”参数，数据类型为“String”，表示语音内容。

语音分割

```
{  
    "source": "s3://path/to/audiol.wav",  
    "usage": "TRAIN",  
    "annotation": [  
        {  
            "type": "modelarts/audio_segmentation",  
            "property": {  
                "@modelarts:start_time": "00:01:10.123",  
                "@modelarts:end_time": "00:01:15.456",  
  
                "@modelarts:source": "Tom",  
  
                "@modelarts:content": "How are you?"  
            },  
            "annotated-by": "human",  
            "creation-time": "2019-01-23 11:30:30"  
        },  
        {  
            "type": "modelarts/audio_segmentation",  
            "property": {  
                "@modelarts:start_time": "00:01:22.754",  
                "@modelarts:end_time": "00:01:24.145",  
                "@modelarts:source": "Jerry",  
                "@modelarts:content": "I'm fine, thank you."  
            },  
            "annotated-by": "human",  
            "creation-time": "2019-01-23 11:30:30"  
        }  
    ]  
}
```

- “source”、“usage”、“annotation”等参数说明与[图片分类](#)一致，详细说明请参见[表2-2](#)。
- “property”的参数解释如[表2-8](#)所示。

表 2-8 “property” 参数说明

参数名	数据类型	描述
@modelarts:start_time	String	声音的起始时间，格式为“hh:mm:ss.SSS”。其中“hh”表示小时，“mm”表示分钟，“ss”表示秒，“SSS”表示毫秒。
@modelarts:end_time	String	声音的结束时间，格式为“hh:mm:ss.SSS”。其中“hh”表示小时，“mm”表示分钟，“ss”表示秒，“SSS”表示毫秒。
@modelarts:source	String	声音来源。
@modelarts:content	String	声音内容。

2.4 标注数据

2.4.1 图像分类

由于模型训练过程需要大量有标签的图片数据，因此在模型训练之前需对没有标签的图片添加标签。您可以通过手工标注或智能一键标注的方式添加标签，快速完成对图片的标注操作，也可以对已标注图片修改或删除标签进行重新标注。

针对图像分类场景，开始标注前，您需要了解：

- 图片标注支持多标签，即一张图片可添加多个标签。
- 标签名是由中文、大小写字母、数字、中划线或下划线组成，且不超过32位的字符串。

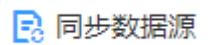
进入数据集详情页

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，基于“标注类型”选择需要进行标注的数据集，单击数据集名称进入数据集详情页。

此操作默认进入数据集当前版本的详情页，如果需要对其他版本进行数据标注，请先在“版本管理”操作中，将需要进行数据标注的版本设置为“当前目录。”详细操作指导请参见[管理数据集版本](#)。

同步数据源

ModelArts会自动从数据集输入位置同步数据至数据集详情页，包含数据及标注信息。

为了快速获取OBS桶中最新数据，可单击数据集详情页中 ，快速将通过OBS上传的数据添加到数据集中。

标注图片（手工标注）

数据集详情页中，展示了此数据集中“未标注”和“已标注”的图片，默认显示“未标注”的图片列表。单击图片右下角 ，即可进行图片的预览，对于已标注图片，预览页面下方会显示该图片的标签信息。

1. 在“未标注”页签，勾选需进行标注的图片。
 - 手工点选：在图片列表中，单击图片，当图片右上角出现蓝色勾选框时，表示已勾选。可勾选同类型的多个图片，一起添加标签。
 - 批量选中：如果图片列表的当前页，所有图片属于一种类型，可以在图片列表的右上角单击“选择当前页”，则当前页面所有的图片将选中。
2. 添加标签。
 - a. 在右侧的“添加标签”区域中，单击“标签名”右侧的文本框中设置标签。
方式一（已存在标签）：单击“标签名”右侧的文本框，然后从下拉列表中选择已有的标签。
方式二（新增标签）：在标签名右侧的文本框中，直接输入新的标签名，然后单击“添加”。
 - b. 查看“选中文件标签”的信息，确认无误后，单击“将选中图片确认为已标注”。此时，选中的图片将被自动移动至“已标注”页签，且在“未标注”页签中，标签的信息也将随着标注步骤进行更新，如增加的标签名称、各标签对应的图片数量。

图 2-4 添加标签



智能标注

除了人工标注外，ModelArts还提供了智能标注功能，快速完成数据标注，为您节省70%以上的标注时间。智能标注是指基于当前标注阶段的标签及图片学习训练，选中系统中已有的模型进行智能标注，快速完成剩余图片的标注操作。

说明

- 启动智能标注时，需数据集存在至少2种标签，且每种标签已标注的图片不少于5张。
 - 启动智能标注时，必须存在未标注图片。
 - 启动智能标注前，保证当前系统中不存在正在进行中的智能标注任务。
1. 在数据集详情页，单击“智能标注”页签，然后单击“启动智能标注”。
 2. 在弹出的“启动智能标注”对话框中，选择智能标注类型，然后单击“提交”。
 - “自动选择”：当系统无可用模型时，可直接设置为自动选择。
 - “自定义选择”：当设置为自定义选择时，必须在选择模型的列表中，选择ModelArts中的已有模型。必须保证ModelArts模型管理列表中，已存在模型，此时，选择模型列表才有数据显示。

图 2-5 启动智能标注

启动智能标注



启动智能标注后，界面显示标注进度。

图 2-6 标注进度

智能标注进度： 70%

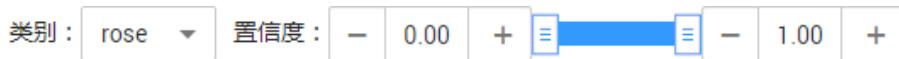
智能标注历史



3. 智能标注完成后，“智能标注”页面将呈现所有标注后的图片列表。通过如下方法确认标注结果。

在图片列表右上角，您可以通过类别或置信度对图片列表进行过滤筛选。

图 2-7 筛选图片



然后，选中一个或多个图片，在右侧的标注信息中查看此图片智能标注结果，如图2-8所示。

- 如果智能添加的标签准确，单击“将选中图片确认为已标注”，此时，选中的图片将被自动移至“已标注”页签，完成数据标注。
- 如果智能添加的标签不准确，您可以单击操作列的编辑按钮，修改标签名称，然后单击“将选中图片确认为已标注”完成标注。

图 2-8 智能标注结果



查看已标注图片

在数据集详情页，单击“已标注”页签，您可以查看已完成标注的图片列表。单击图片，可在右侧的“选中文件标签”中了解当前图片的标签信息。

修改标注

当数据完成标注后，您还可以进入已标注页签，对已标注的数据进行修改。

● 基于图片修改

在数据标注页面，单击“已标注”页签，然后在图片列表中选中待修改的图片（选择一个或多个）。在右侧标签信息区域中对图片信息进行修改。

- 添加标签：在“标签名”右侧文本框中，选择已有标签或输入新的标签名，然后单击“确定”，为选中图片增加标签。
- 修改标签：在“选中文件标签”区域中，单击操作列的编辑图标，然后在文本框中输入正确的标签名，然后单击完成修改。

图 2-9 编辑标签



- 删除标签：在“选中文件标签”区域中，单击操作列的 删除该标签。

- 基于标签修改

在数据标注页面，单击“已标注”页签，在图片列表右侧，显示全部标签的信息。

图 2-10 全部标签的信息

The screenshot shows a table titled '全部标签' (All Labels) with three columns: '标签' (Label), '数量' (Quantity), and '操作' (Operation). It lists three labels: 'sunflower' (9), 'daisy' (11), and 'rose' (11), each with an edit icon and a delete icon.

标签	数量	操作
sunflower	9	✎ ⚡
daisy	11	✎ ⚡
rose	11	✎ ⚡

- 修改标签：单击操作列的 ，然后在弹出的对话框中输入修改后的标签名，然后单击“确定”完成修改。修改后，之前添加了此标签的图片，都将被标注为新的标签名称。
- 删除标签：单击操作列的 ，在弹出的对话框中，选择“仅删除标签”或“删除标签及仅包含此标签的图片”，然后单击“确定”。

图 2-11 删除标签



添加图片

除了数据集输入位置自动同步的数据外，您还可以在ModelArts界面中，直接添加图片，用于数据标注。

1. 在数据集详情页面，单击“未标注”页签，然后单击左上角“添加图片”。

2. 在弹出的“添加图片”对话框中，单击“添加图片”。

选择本地环境中需要上传的图片，可以一次性选择多张图片。图片只支持JPG、JPEG、PNG、BMP格式，且一次上传图片的总大小不能超过8MB。

图片选择完成后，“添加图片”对话框将显示上传图片的缩略图以及图片大小。

图 2-12 添加图片



3. 在添加图片对话框中，单击“确定”，完成添加图片的操作。

您添加的图片将自动呈现在“未标注”的图片列表中。且图片将自动存储至此“数据集输入位置”对应的OBS目录中。

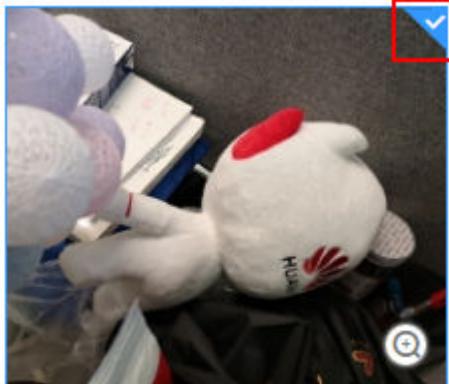
删除图片

通过数据删除操作，可将需要丢弃的图片数据快速删除。

在“未标注”或“已标注”页面中，依次单击选中需要删除的图片，或者选择“勾选当前页”选中该页面所有图片，然后单击左上角“删除图片”，即可完成图片的删除操作。

其中，被选中的图片，其右上角将显示为勾选状态。如果当前页面无选中图片时，“删除图片”按钮为灰色，无法执行删除操作。

图 2-13 选中图片



注意

删除图片操作是将删除对应OBS目录下存储的图片。删除后，数据将无法恢复，请谨慎操作。

2.4.2 物体检测

由于模型训练过程需要大量有标签的图片数据，因此在模型训练之前需对没有标签的图片添加标签。您可以通过手工标注或智能一键标注的方式添加标签，快速完成对图片的标注操作，也可以对已标注图片修改或删除标签进行重新标注。

针对物体检测场景，开始标注前，您需要了解：

- 图片中所有目标物体都要标注。
- 目标物体清晰无遮挡的，必须画框。
- 画框仅包含整个物体。框内包含整个物体的全部，画框边缘不要压着物体，边缘和物体间不要留着空隙，避免背景对模型训练造成干扰。

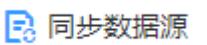
进入数据集详情页

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，基于“标注类型”选择需要进行标注的数据集，单击数据集名称进入数据集详情页。

此操作默认进入数据集当前版本的详情页，如果需要对其他版本进行数据标注，请先在“版本管理”操作中，将需要进行数据标注的版本设置为“当前目录。”
详细操作指导请参见[管理数据集版本](#)。

同步数据源

ModelArts会自动从数据集输入位置同步数据至数据集详情页，包含数据及标注信息。

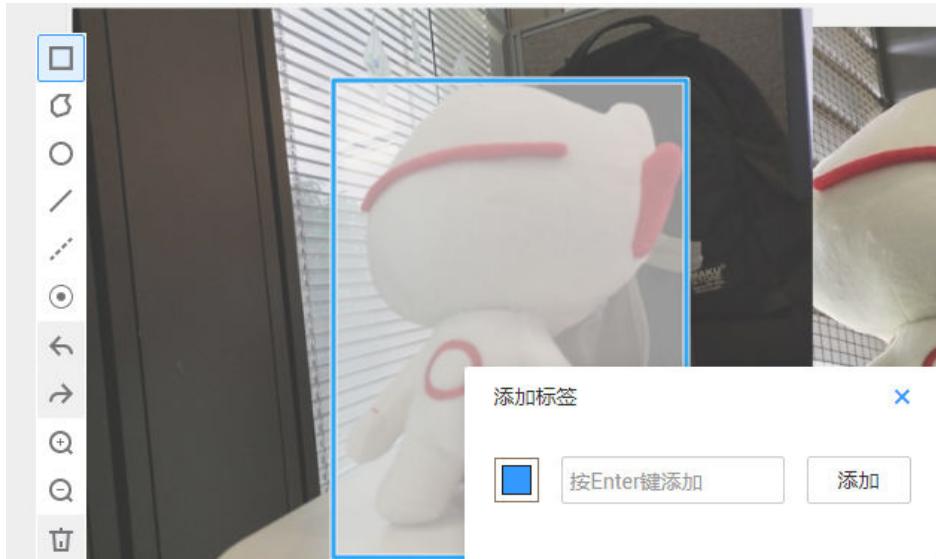
为了快速获取OBS桶中最新数据，可单击数据集详情页中 ，快速将通过OBS上传的数据添加到数据集中。

标注图片（手工标注）

数据集详情页中，展示了此数据集中“未标注”和“已标注”的图片，默认显示“未标注”的图片列表。

1. 在“未标注”页签图片列表中，单击图片，自动跳转到标注页面。
 2. 在页面左侧工具栏选择合适的标注图形，该样例单击选择“”框选图片进行标注。
-  **说明**
- 页面左侧可以选择多种形状对图片进行标注。标注第一张图片时，一旦选择其中一种，其他所有图片都需要使用此形状进行标注。
3. 在弹出的添加标签文本框中，直接输入新的标签名，然后单击“添加”。如果已存在标签，从下拉列表中选择已有的标签，单击“添加”。

图 2-14 添加标签



4. 单击页面上方“返回数据标注预览”查看标注信息，在弹框中单击“确定”保存当前标注并离开标注页面。
5. 选中的图片被自动移动至“已标注”页签，且在“未标注”页签中，标签的信息也将随着标注步骤进行更新，如增加的标签名称、标签对应的图片数量。

智能标注

除了人工标注外，ModelArts还提供了智能标注功能，快速完成数据标注，为您节省70%以上的标注时间。智能标注是指基于当前标注阶段的标签及图片学习训练，选中系统中已有的模型进行智能标注，快速完成剩余图片的标注操作。



- 启动智能标注时，必须存在未标注图片。
- 启动智能标注前，保证当前系统中不存在正在进行中的智能标注任务。

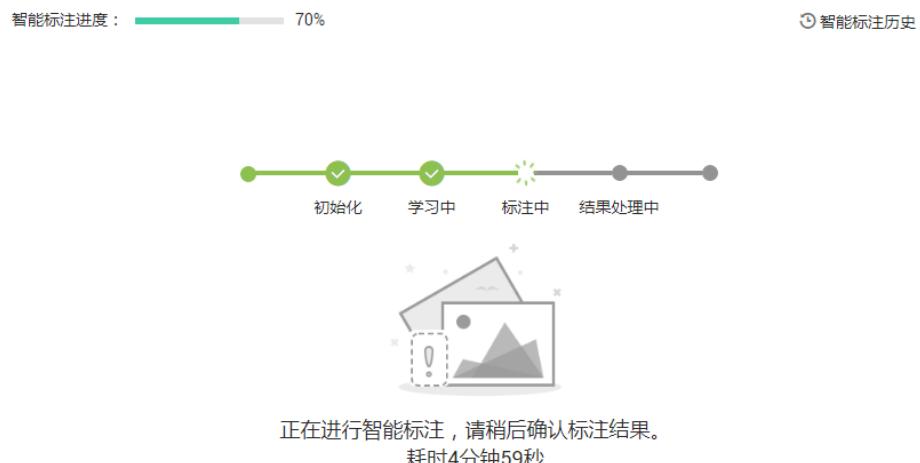
1. 在数据集详情页，单击“智能标注”页签，然后单击“启动智能标注”。
2. 在弹出的“启动智能标注”对话框中，选择智能标注类型，然后单击“提交”。
 - “自动选择”：当系统无可用模型时，可直接设置为自动选择。
 - “自定义选择”：当设置为自定义选择时，必须在选择模型的列表中，选择 ModelArts 中的已有模型。必须保证 ModelArts 模型管理列表中，已存在模型，此时，选择模型列表才有数据显示。

图 2-15 启动智能标注



启动智能标注后，界面显示标注进度。

图 2-16 标注进度



3. 智能标注完成后，“智能标注”页面将呈现所有标注后的图片列表。通过如下方法确认标注结果。
在图片列表右上角，您可以通过类别或置信度对图片列表进行过滤筛选。

图 2-17 筛选图片

The screenshot shows the image filtering interface. It includes a 'Category' filter (类别) set to 'rose' and a 'Confidence Level' filter (置信度) with a range from 0.00 to 1.00. The confidence level slider is currently positioned between 0.00 and 1.00.

然后，选中一个图片，自动跳转到标注页面。在右侧的标注信息中查看此图片智能标注结果。

- 如果智能添加的标签准确，单击“确认为已标注”，此时，选中的图片将被自动移至“已标注”页签，完成数据标注。
- 如果智能添加的标签不准确，您可以单击操作列的编辑按钮，修改标签名称，然后单击“确认为已标注”完成标注。

查看已标注图片

在数据集详情页，单击“已标注”页签，您可以查看已完成标注的图片列表。单击图片，可在右侧的“当前文件标签”中了解当前图片的标签信息。

修改标注

当数据完成标注后，您还可以进入已标注页签，对已标注的数据进行修改。

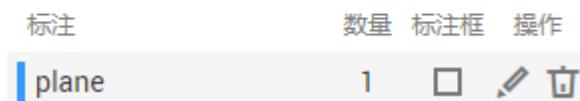
● 基于图片修改

在数据标注页面，单击“已标注”页签，然后在图片列表中选中待修改的图片。自动跳转到标注页面，在右侧标签信息区域中对图片信息进行修改。

- 修改标签：“当前文件标签”区域中，单击操作列的，然后在文本框中输入正确的标签名，然后单击完成修改。
- 删除标签：在“当前文件标签”区域中，单击操作列的删除该标签，标签删除后，单击页面上方“返回数据标注预览”，在弹框中单击“确定”保存当前操作并离开标注页面。该图片会重新回到“未标注”页签。

图 2-18 编辑标签

当前文件标签



● 基于标签修改

在数据标注页面，单击“已标注”页签，在图片列表右侧，显示全部标签的信息。

图 2-19 全部标签的信息

全部标签		
标签	数量	操作
sunflower	9	 
daisy	11	 
rose	11	 

- 修改标签：单击操作列的 ，然后在弹出的对话框中输入修改后的标签名，然后单击“确定”完成修改。修改后，之前添加了此标签的图片，都将被标注为新的标签名称。
- 删除标签：单击操作列的 ，在弹出的对话框中，选择“仅删除标签”或“删除标签及仅包含此标签的图片”，然后单击“确定”。

添加图片

除了数据集输入位置自动同步的数据外，您还可以在ModelArts界面中，直接添加图片，用于数据标注。

1. 在数据集详情页面，单击“未标注”页签，然后单击左上角“添加图片”。
2. 在弹出的“添加图片”对话框中，单击“添加图片”。
选择本地环境中需要上传的图片，可以一次性选择多张图片。图片只支持JPG、JPEG、PNG、BMP格式，且一次上传图片的总大小不能超过8MB。
图片选择完成后，“添加图片”对话框将显示上传图片的缩略图以及图片大小。

图 2-20 添加图片



3. 在添加图片对话框中，单击“确定”，完成添加图片的操作。
您添加的图片将自动呈现在“未标注”的图片列表中。且图片将自动存储至此“数据集输入位置”对应的OBS目录中。

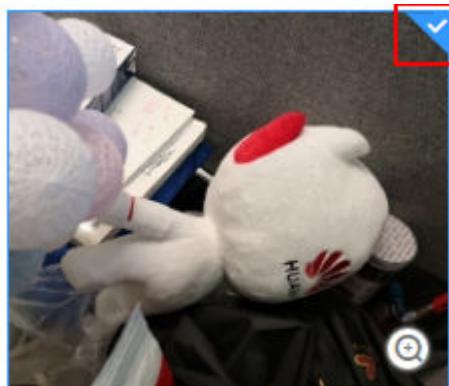
删除图片

通过数据删除操作，可将需要丢弃的图片数据快速删除。

在“未标注”或“已标注”页面中，依次单击选中需要删除的图片，或者选择“勾选当前页”选中该页面所有图片，然后单击左上角“删除图片”，即可完成图片的删除操作。

其中，被选中的图片，其右上角将显示为勾选状态。如果当前页面无选中图片时，“删除图片”按钮为灰色，无法执行删除操作。

图 2-21 选中图片



注意

删除图片操作是将删除对应OBS目录下存储的图片。删除后，数据将无法恢复，请谨慎操作。

2.4.3 文本分类

由于模型训练过程需要大量有标签的数据，因此在模型训练之前需对没有标签的文本添加标签。您也可以对已标注文本进行修改、删除和重新标注。

针对文本分类场景，是对文本的内容按照标签进行分类处理，开始标注前，您需要了解：

- 文本标注支持多标签，即一个标注对象可添加多个标签。
- 标签名是由中文、大小写字母、数字、中划线或下划线组成，且不超过32位的字符串。

进入数据集详情页

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，基于“标注类型”选择需要进行标注的数据集，单击数据集名称进入数据集详情页。

此操作默认进入数据集当前版本的详情页，如果需要对其他版本进行数据标注，请先在“版本管理”操作中，将需要进行数据标注的版本设置为“当前目录。”详细操作指导请参见[管理数据集版本](#)。

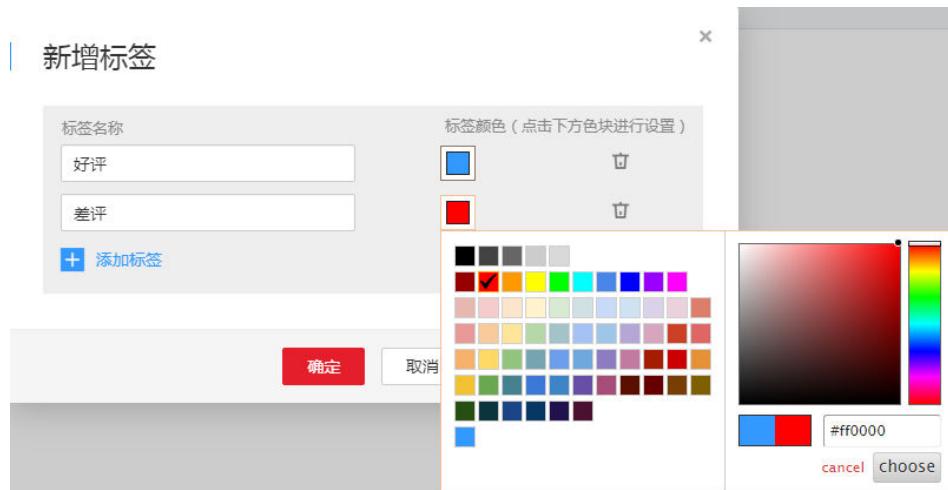
标注文本

数据集详情页中，展示了此数据集中“未标注”和“已标注”的文本，默认显示“未标注”的文本列表。

1. 添加标签。

- 在“未标注”页签添加：在右侧单击页面中标签集 ，在“新增标签”页中添加标签名称，选择标签颜色，单击“确定”完成标签的新增。
- 在“已标注”页签添加：在右侧单击页面中全部标签 ，在“新增标签”页中添加标签名称，选择标签颜色，单击“确定”完成标签的新增。

图 2-22 新增文本标签



2. 在“未标注”页签文本列表中，单击文本，当文本背景变为蓝色时，表示已选择。选择上方标签集中不同标签进行标注。一个标注对象可添加多个标签。

图 2-23 文本分类标注



3. 单击页面下方“保存当前页”完成文本标注。

查看已标注文本

在数据集详情页，单击“已标注”页签，您可以查看已完成标注的文本列表。单击文本框，可在右侧的“全部标签”中了解全部已标注文本的标签信息。

修改标注

当数据完成标注后，您还可以进入已标注页签，对已标注的数据进行修改。

● 基于文本修改

在数据集详情页，单击“已标注”页签，然后在文本列表中选中待修改的文本。

- 手工点选删除：在文本列表中，单击文本，当文本背景变为蓝色时，表示已选择。当文本有多个标签时，可以单击文本标签上方的~~×~~删除单个标签。您也可以选中文本后，单击页面上方“删除”按钮，删除当前文本所有标签，该文本会重新回到“未标注”页签。
- 批量选中删除：在文本列表的右上角单击“选择当前页”，则当前页面所有的文本将选中，单击页面上方“删除”按钮，删除本页文本所有标签。

● 基于标签修改

在数据集详情页，单击“已标注”页签，在图片列表右侧，显示全部标签的信息。

- 批量修改：在“全部标签”区域中，单击操作列的笔图标，然后在文本框中添加标签名称，选择标签颜色，单击“确定”完成修改。
- 批量删除：在“全部标签”区域中，单击操作列的~~垃圾桶图标~~删除该标签，在弹出对话框中，可选择“仅删除标签”或“删除标签及仅包含此标签的标注对象”，然后单击“确定”。

添加文件

除了数据集输入位置自动同步的数据外，您还可以在ModelArts界面中，直接添加文件，用于数据标注。

1. 在数据集详情页面，单击“未标注”页签，然后单击左上角“添加文件”。
2. 在弹出的“添加文件”对话框中，选择上传文件。

选择本地环境中需要上传的文件，可以一次性选择多个文件。文件格式只支持txt或csv，且一次上传文件的总大小不能超过8MB。

图 2-24 添加图片



3. 在添加文件对话框中，单击“确定”，完成添加文件的操作。您添加的文件内容将自动呈现在“未标注”的文本列表中。

删除文件

通过数据删除操作，可将需要丢弃的文件数据快速删除。

- 在“未标注”页面中，单击选中需要删除的文本，然后单击左上角“删除”，即可完成文本的删除操作。
- 在“已标注”页面中，单击依次选中需要删除的文本，或者选择“勾选当前页”选中该页面所有文本，然后单击左上角“删除”，即可完成文本的删除操作。

其中，被选中的文本，其背景将显示为蓝色。如果当前页面无选中文本时，“删除图片”按钮为灰色，无法执行删除操作。

2.4.4 命名实体

命名实体场景，是针对文本中的实体片段进行标注，如“时间”、“地点”等。开始标注前，您需要了解：

- 实体命名标签名是由中文、大小写字母、数字、中划线或下划线组成，且不超过32位的字符串。

进入数据集详情页

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，基于“标注类型”选择需要进行标注的数据集，单击数据集名称进入数据集详情页。

此操作默认进入数据集当前版本的详情页，如果需要对其他版本进行数据标注，请先在“版本管理”操作中，将需要进行数据标注的版本设置为“当前目录。”详细操作指导请参见[管理数据集版本](#)。

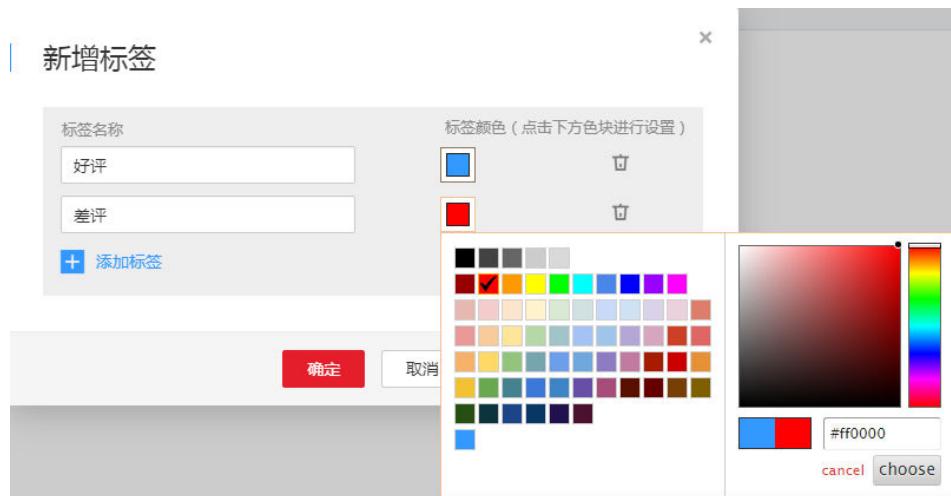
标注文本

数据集详情页中，展示了此数据集中“未标注”和“已标注”的文本，默认显示“未标注”的文本列表。

1. 添加标签。

- 在“未标注”页签添加：在右侧单击页面中标签集，在“新增标签”页中添加标签名称，选择标签颜色，单击“确定”完成标签的新增。
- 在“已标注”页签添加：在右侧单击页面中全部标签，在“新增标签”页中添加标签名称，选择标签颜色，单击“确定”完成标签的新增。

图 2-25 新增文本标签



2. 在“未标注”页签文本列表中，单击文本，鼠标选中待标注文本内容，然后点击标签集区域中不同标签进行标注。

图 2-26 命名实体标注



3. 单击页面下方“保存当前页”完成文本标注。

查看已标注文本

在数据集详情页，单击“已标注”页签，您可以查看已完成标注的文本列表。单击文本框，可在右侧的“全部标签”中了解全部已标注文本的标签信息。

修改标注

当数据完成标注后，您还可以进入“已标注”页签，对已标注的数据进行修改。

在数据集详情页，单击“已标注”页签，在右侧标签信息区域中对文本信息进行修改。

● 基于文本修改

在数据集详情页，单击“已标注”页签，然后在文本列表中选中待修改的文本。

- 手工点选删除：在文本列表中，单击文本，当文本背景变为蓝色时，表示已选择。在页面右侧，单击文本标签上方的‘X’删除单个标签。您也可以选中文本后，单击页面上方“删除”按钮，删除当前文本所有标签，该文本会重新回到“未标注”页签。
- 批量选中删除：在文本列表上方勾选“选择当前页”，则当前页面所有的文本将选中，单击页面上方“删除”按钮，删除本页文本所有标签。

● 基于标签修改

在数据集详情页，单击“已标注”页签，在图片列表右侧，显示全部标签的信息。

- 批量修改：在“全部标签”区域中，单击操作列的，然后在文本框中添加标签名称，选择标签颜色，单击“确定”完成修改。
- 批量删除：在“全部标签”区域中，单击操作列的删除该标签，在弹出对话框中，可选择“仅删除标签”或“删除标签及仅包含此标签的标注对象”，然后单击“确定”。

添加文件

除了数据集输入位置自动同步的数据外，您还可以在ModelArts界面中，直接添加文件，用于数据标注。

1. 在数据集详情页面，单击“未标注”页签，然后单击左上角“添加文件”。
2. 在弹出的“添加文件”对话框中，选择上传文件。

选择本地环境中需要上传的文件，可以一次性选择多个文件。文件格式只支持txt或csv，且一次上传文件的总大小不能超过8MB。

图 2-27 添加图片



3. 在添加文件对话框中，单击“确定”，完成添加文件的操作。您添加的文件内容将自动呈现在“未标注”的文本列表中。

删除文件

通过数据删除操作，可将需要丢弃的文件数据快速删除。

- 在“未标注”页面中，单击选中需要删除的文本，然后单击左上角“删除”，即可完成文本的删除操作。
- 在“已标注”页面中，单击依次选中需要删除的文本，或者选择“勾选当前页”选中该页面所有文本，然后单击左上角“删除”，即可完成文本的删除操作。

其中，被选中的文本，其背景将显示为蓝色。如果当前页面无选中文本时，“删除图片”按钮为灰色，无法执行删除操作。

2.4.5 声音分类

由于模型训练过程需要大量有标签的音频数据，因此在模型训练之前需对没有标签的音频添加标签。通过ModelArts您可对音频进行一键式批量添加标签，快速完成对音频的标注操作，也可以对已标注音频修改或删除标签进行重新标注。

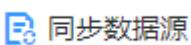
进入数据集详情页

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，基于“标注类型”选择需要进行标注的数据集，单击数据集名称进入数据集详情页。

此操作默认进入数据集当前版本的详情页，如果需要对其他版本进行数据标注，请先在“版本管理”操作中，将需要进行数据标注的版本设置为“当前目录。”详细操作指导请参见[管理数据集版本](#)。

同步数据源

ModelArts会自动从数据集输入位置同步数据至数据集详情页，包含数据及标注信息。

为了快速获取OBS桶中最新数据，可单击数据集详情页中，快速将通过OBS上传的数据添加到数据集中。

标注音频

数据集详情页中，展示了此数据集中“未标注”和“已标注”的音频，默认显示“未标注”的音频列表。单击音频左侧，即可进行音频的试听。

1. 在“未标注”页签，勾选需进行标注的音频。
 - 手工点选：在音频列表中，单击音频，当右上角出现蓝色勾选框时，表示已勾选。可勾选同类别的多个音频，一起添加标签。
 - 批量选中：如果音频列表的当前页，所有音频属于一种类型，可以在列表的右上角单击“选择当前页”，则当前页面所有的音频将选中。
2. 添加标签。
 - a. 在右侧的“标签”区域中，单击“标签名”右侧的文本框中设置标签。
方式一（已存在标签）：单击“标签名”右侧的文本框，然后从下拉列表中选择已有的标签。
方式二（新增标签）：在标签名右侧的文本框中，直接输入新的标签名，然后单击“确定”。
 - b. 选中的音频将被自动移动至“已标注”页签，且在“未标注”页签中，标签的信息也将随着标注步骤进行更新，如增加的标签名称、各标签对应的音频数量。

图 2-28 添加标签



查看已标注音频

在数据集详情页，单击“已标注”页签，您可以查看已完成标注的音频列表。单击音频，可在右侧的“选中文件标签”中了解当前音频的标签信息。

修改标注

当数据完成标注后，您还可以进入“已标注”页签，对已标注的数据进行修改。

● 基于音频修改

在数据标注页面，单击“已标注”页签，然后在音频列表中选中待修改的音频（选择一个或多个）。在右侧标签信息区域中对标签进行修改。

- 修改标签：在“选中文件标签”区域中，单击操作列的，然后在文本框中输入正确的标签名，然后单击完成修改。

图 2-29 编辑标签



- 删除标签：在“选中文件标签”区域中，单击操作列的删除该标签。

● 基于标签修改

在数据标注页面，单击“已标注”页签，在音频列表右侧，显示全部标签的信息。

图 2-30 全部标签的信息

全部标签 1			
标签	数量	快捷键	操作
aa	23	1	

- 修改标签：单击操作列的，然后在弹出的对话框中输入修改后的标签名，然后单击“确定”完成修改。修改后，之前添加了此标签的音频，都将被标注为新的标签名称。
- 删除标签：单击操作列的，在弹出的对话框中，选择“仅删除标签”或“删除标签及仅包含此标签的图片”，然后单击“确定”。

添加音频

除了数据集输入位置自动同步的数据外，您还可以在ModelArts界面中，直接添加音频，用于数据标注。

1. 在数据集详情页面，单击“未标注”页签，然后单击左上角“添加音频”。
2. 在弹出的“添加音频”对话框中，单击“添加音频”。

选择本地环境中需要上传的音频，仅支持WAV格式音频文件，单个音频文件不能超过4MB，且单次上传的音频文件总大小不能超过8MB。

- 在添加音频对话框中，单击“确定”，完成添加音频的操作。

您添加的音频将自动呈现在“未标注”的音频列表中。且音频将自动存储至此“数据集输入位置”对应的OBS目录中。

删除音频

通过数据删除操作，可将需要丢弃的音频数据快速删除。

在“未标注”或“已标注”页面中，依次单击选中需要删除的音频，或者选择“勾选当前页”选中该页面所有音频，然后单击左上角“删除音频”，即可完成音频的删除操作。

其中，被选中的音频，其右上角将显示为勾选状态。如果当前页面无选中音频时，“删除音频”按钮为灰色，无法执行删除操作。

注意

删除音频操作是将删除对应OBS目录下存储的音频。删除后，数据将无法恢复，请谨慎操作。

2.4.6 语音内容

由于模型训练过程需要大量有标签的音频数据，因此在模型训练之前需对没有标签的音频添加标签。通过ModelArts您可对音频进行一键式批量添加标签，快速完成对音频的标注操作，也可以对已标注音频修改或删除标签进行重新标注。

进入数据集详情页

- 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
- 在数据集列表中，基于“标注类型”选择需要进行标注的数据集，单击数据集名称进入数据集详情页。
此操作默认进入数据集当前版本的详情页，如果需要对其他版本进行数据标注，请先在“版本管理”操作中，将需要进行数据标注的版本设置为“当前目录。”
详细操作指导请参见[管理数据集版本](#)。

同步数据源

ModelArts会自动从数据集输入位置同步数据至数据集详情页，包含数据及标注信息。

同步数据源

为了快速获取OBS桶中最新数据，可单击数据集详情页中[同步数据源](#)，快速将通过OBS上传的数据添加到数据集中。

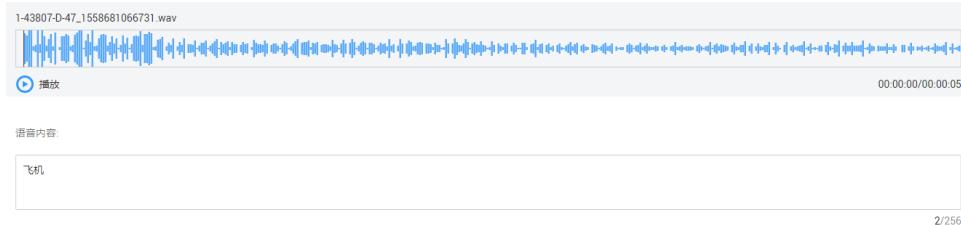
标注音频

数据集详情页中，展示了此数据集中“未标注”和“已标注”的音频，默认显示“未标注”的音频列表。

- 在“未标注”页签左侧音频列表中，单击目标音频文件，在右侧的区域中出现音频，单击音频下方，即可进行音频播放。

2. 根据播放内容，在下方“语音内容”文本框中填写音频内容。
3. 输入内容后单击下方的“确认标注”按钮完成标注。音频将被自动移动至“已标注”页签，且在“未标注”页签中，标签的信息也将随着标注步骤进行更新，如各标签对应的音频数量。

图 2-31 语音内容音频标注



查看已标注音频

在数据集详情页，单击“已标注”页签，您可以查看已完成标注的音频列表。单击音频，可在右侧的“语音内容”文本框中了解当前音频的内容信息。

修改标注

当数据完成标注后，您还可以进入“已标注”页签，对已标注的数据进行修改。

在数据集详情页，单击“已标注”页签，然后在音频列表中选中待修改的音频。在右侧标签信息区域中修改“语音内容”文本框中的内容，单击下方的“确认标注”按钮完成修改。

添加音频

除了数据集输入位置自动同步的数据外，您还可以在ModelArts界面中，直接添加音频，用于数据标注。

1. 在数据集详情页面，单击“未标注”页签，然后单击左上角“添加音频”。
2. 在弹出的“添加音频”对话框中，单击“添加音频”。
选择本地环境中需要上传的音频，仅支持WAV格式音频文件，单个音频文件不能超过4MB，且单次上传的音频文件总大小不能超过8MB。
3. 在添加音频对话框中，单击“确定”，完成添加音频的操作。
您添加的音频将自动呈现在“未标注”的音频列表中。且音频将自动存储至此“数据集输入位置”对应的OBS目录中。

删除音频

通过数据删除操作，可将需要丢弃的音频数据快速删除。

在“未标注”或“已标注”页面中，单击选中需要删除的音频，然后单击左上角“删除音频”，即可完成音频的删除操作。

如果当前页面无选中音频时，“删除音频”按钮为灰色，无法执行删除操作。

注意

删除音频操作是将删除对应OBS目录下存储的音频。删除后，数据将无法恢复，请谨慎操作。

2.4.7 语音分割

由于模型训练过程需要大量有标签的音频数据，因此在模型训练之前需对没有标签的音频添加标签。通过ModelArts您可对音频添加标签，快速完成对音频的标注操作，也可以对已标注音频修改或删除标签进行重新标注。

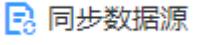
进入数据集详情页

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，基于“标注类型”选择需要进行标注的数据集，单击数据集名称进入数据集详情页。

此操作默认进入数据集当前版本的详情页，如果需要对其他版本进行数据标注，请先在“版本管理”操作中，将需要进行数据标注的版本设置为“当前目录。”详细操作指导请参见[管理数据集版本](#)。

同步数据源

ModelArts会自动从数据集输入位置同步数据至数据集详情页，包含数据及标注信息。

为了快速获取OBS桶中最新数据，可单击数据集详情页中  同步数据源，快速将通过OBS上传的数据添加到数据集中。

标注音频

数据集详情页中，展示了此数据集中“未标注”和“已标注”的音频，默认显示“未标注”的音频列表。

1. 在“未标注”页签左侧音频列表中，单击目标音频文件，在右侧的区域中出现音频，单击音频下方 ，即可进行音频播放。
2. 根据播放内容，选取合适的音频段，在下方“语音内容”文本框红填写音频标签和内容。

图 2-32 语音标签音频标注



3. 输入内容后单击下方的“确认标注”按钮完成标注。音频将被自动移动至“已标注”页签，且在“未标注”页签中，标签的信息也将随着标注步骤进行更新，如各标签对应的音频数量。

查看已标注音频

在数据集详情页，单击“已标注”页签，您可以查看已完成标注的音频列表。单击音频，可在右侧的“语音内容”文本框中了解当前音频的内容信息。

修改标注

当数据完成标注后，您还可以进入“已标注”页签，对已标注的数据进行修改。

- 修改标签：在数据集详情页，单击“已标注”页签，然后在音频列表中选中待修改的音频。在右侧标签信息区域中修改“语音内容”中的“标签”和“内容”，单击下方的“确认标注”按钮完成修改。
- 删除标签：单击目标编号操作列的 ，删除该段音频的标注。您也可以单击标注音频文件上方的 删除标注，然后单击“确认标注”。

添加音频

除了数据集输入位置自动同步的数据外，您还可以在ModelArts界面中，直接添加音频，用于数据标注。

1. 在数据集详情页面，单击“未标注”页签，然后单击左上角“添加音频”。
2. 在弹出的“添加音频”对话框中，单击“添加音频”。
选择本地环境中需要上传的音频，仅支持WAV格式音频文件，单个音频文件不能超过4MB，且单次上传的音频文件总大小不能超过8MB。
3. 在添加音频对话框中，单击“确定”，完成添加音频的操作。
您添加的音频将自动呈现在“未标注”的音频列表中。且音频将自动存储至此“数据集输入位置”对应的OBS目录中。

删除音频

通过数据删除操作，可将需要丢弃的音频数据快速删除。

在“未标注”或“已标注”页面中，单击选中需要删除的音频，然后单击左上角“删除音频”，即可完成音频的删除操作。

如果当前页面无选中音频时，“删除音频”按钮为灰色，无法执行删除操作。

注意

删除音频操作是将删除对应OBS目录下存储的音频。删除后，数据将无法恢复，请谨慎操作。

2.5 发布数据集

ModelArts在数据集管理过程中，针对同一个数据源，对不同时间标注后的数据，按版本进行区分，方便后续模型构建和开发过程中，选择对应的数据集版本进行使用。数据标注完成后，您可以将数据集当前状态进行发布，生成一个新的数据集版本。

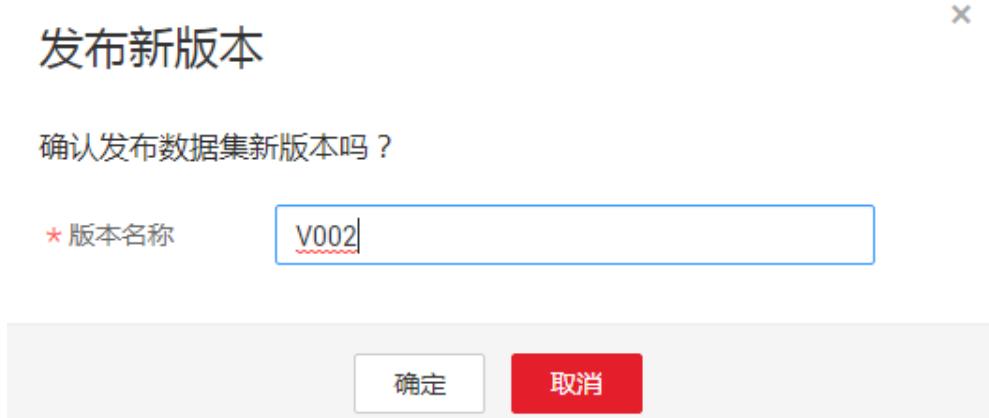
关于数据集版本

- 针对刚创建的数据集（未发布前），无数据集版本信息，必须执行发布操作后，才能应用于模型开发或训练。
- 数据集版本，默认按V001、V002规则进行命名，您也可以在发布时自定义设置。
- 您可以将任意一个版本设置为当前目录，即表示数据集列表中进入的数据集详情，为此版本的数据及标注信息。

发布数据集

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，单击操作列的“发布”，在“发布新版本”弹出框中，填写“版本名称”，然后单击“确定”。
版本名称默认按V001、V002规则进行命名，您也可以设置为自定义的名称。名称只能包含字母、数字、中划线或下划线。

图 2-33 发布数据集



版本发布后，您可以前往版本管理查看详细信息。系统默认将最新的版本作为当前目录。

数据集发布后，相关文件的目录结构说明

由于数据集是基于OBS目录管理的，发布为新版本后，对应的数据集输出位置，也将基于新版本生成目录。

以图像分类为例，数据集发布后，对应OBS路径下生成，其相关文件的目录如下所示：

```
|-- user-specified-output-path  
|   |-- DatasetName-datasetId
```

```
|-- annotation
    |-- VersionMamel
        |-- VersionMamel.manifest
    |-- VersionName2
        ...
    |-- ...
```

如果数据集导入的是Manifest文件，在数据集发布后，其相关文件的目录结构如下。

```
|-- user-specified-output-path  
    |-- DatasetName-datasetId  
        |-- annotation  
            |-- VersionMame1  
                |-- VersionMame1.manifest  
                |-- file1.jpg  
                |-- file1.xml  
                |-- ...  
            |-- VersionMame2  
                ...  
            |-- ...
```

2.6 管理数据集版本

数据标注完成后，您可以发布成多个版本对数据集进行管理。针对已发布生产的数据集版本，您可以通过查看数据集演进过程、设置当前版本、删除版本等操作，对数据集进行管理。数据集版本的相关说明，请参见[关于数据集版本](#)。

发布为新版本的说明，请参见[发布数据集](#)。

查看数据集演进过程

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
 2. 在数据集列表中，单击操作列的“版本管理”，进入数据集版本管理页面。

您可以查看数据集版本的“名称”、“状态”、“文件总数”、“已标注文件个数”，并在左侧的“演讲过程”中查看版本的发布时间。

图 2-34 查看数据集版本

演进过程	版本名称	状态	文件总数	已标注文件个数	操作
2019/05/27 16:33....	V002 当前目录	正常	3658	27	删除
2019/05/27 16:05....	V001	正常	3658	11	删除

设置当前版本

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
 2. 在数据集列表中，单击操作列的“版本管理”，进入数据集版本管理页面。
 3. 在版本管理页面中，将鼠标移至版本所在行，当出现“设置为当前目录”时，单击此按钮。设置完成后，版本名称右侧将显示为“当前目录”。



只有状态为“正常”的版本，才能被设置为当前目录。

图 2-35 设置当前版本

演进过程	版本名称	状态
2019/05/27 16:33:...	V002 设置为当前目录	✓ 正常
2019/05/27 16:05:...	V001 当前目录	✓ 正常

删除数据集版本

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，单击操作列的“版本管理”，进入数据集版本管理页面。
3. 选择需删除的版本所在行，单击操作列的“删除”。在弹出的对话框中确认信息，然后单击“确定”完成删除操作。

说明

删除数据集版本不会删除原始数据，数据及其标注信息仍存在在对应的OBS目录下。但是，执行删除操作后，无法在ModelArts管理控制台清晰的管理数据集版本，请谨慎操作。

2.7 修改数据集

对于已创建的数据集，您可以修改数据集的基本信息以匹配业务变化。

前提条件

已存在创建完成的数据集。

修改数据集基本信息

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据标注”，进入“数据集”管理页面。
2. 在数据集列表中，单击操作列的“修改”，参考表2-9修改数据集基本信息，然后单击“确定”完成修改。

图 2-36 修改数据集

修改数据集

名称	dataset-mmst								
描述	0/256								
标签集	<table border="1"><tr><td>标签名称</td><td>1</td><td>标签颜色</td><td>②</td></tr><tr><td colspan="4">+ 添加标签</td></tr></table>	标签名称	1	标签颜色	②	+ 添加标签			
标签名称	1	标签颜色	②						
+ 添加标签									

表 2-9 参数说明

参数	说明
名称	数据集的名称，名称只能是字母、数字、下划线、中划线或者中文字符组成的合法字符串。
描述	数据集的简要描述。
标签集	<p>仅“图像分类”、“物体检测”、“声音分类”、“命名实体”、“文本分类”类型的数据集涉及修改“标签集”，且只有“命名实体”、“物体检测”、“文本分类”类型的数据集涉及修改“标签颜色”。</p> <ul style="list-style-type: none">● 修改标签名称：在标签名称文本框中，修改标签名称。标签名称只能是字母、数字、下划线或中划线组成的合法字符串，不支持中文。长度为1~32字符。● 添加标签：单击  添加标签 添加标签。● 设置标签颜色：在每个标签右侧的标签颜色区域下， 

2.8 删除数据集

如果数据集不再使用，您可以删除数据集释放资源。

说明

删除数据集后，数据集对应的数据集输入位置和数据集输出位置对应的OBS目录下，如果需要删除OBS目录下的数据释放资源，建议前往OBS管理控制台，删除对应的数据，然后再删除OBS文件夹。

操作步骤

1. 在“数据标注”页面中，单击数据集操作列的“删除”。
2. 在弹出的对话框中，单击“确定”，确认删除此数据集。

说明

删除后，数据集的版本管理等功能无法恢复，请谨慎操作。但是，此数据集对应的原始数据和标注数据依然存储在OBS中。

3 数据管理

3.1 数据管理简介

在使用数据进行模型训练之前，您可在“数据管理”页面对数据进行处理、创建数据集并进行数据集版本管理。

- **数据标注**

可创建数据标注作业，可进行图像分类、物体检测两种类型的人工标注。

- **数据集**

可创建数据集、发布多个版本，并可进行版本之间的对比。

数据迁移

由于数据标注即将下线，推荐使用数据标注（Beta），如需将数据迁移至数据标注（Beta），请单击一键迁移按钮进行迁移。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理”。
2. 在数据标注页面，单击一键迁移，将已有的数据标注作业迁移至数据标注模块。

 **说明**

- 当前仅支持迁移数据标注作业。针对数据集，其数据依然存储在用户的OBS桶内，建议直接在数据标注模块，直接使用此目录创建数据集即可。
- 对于正在迁移的数据标注作业，请勿执行其他操作。

3.2 数据标注

3.2.1 数据标注界面说明

在数据管理“数据标注”页面列举了用户所创建的数据标注作业。在该界面您可以[创建数据标注作业](#)，也可以在列表右上方搜索框中输入名称，单击进行查询，数据标注列表说明如[表3-1](#)所示。

表 3-1 数据标注列表说明

参数名称	说明
名称	数据标注作业的名称。单击名称，进入该标注作业详情页面。
标注类型	当前数据标注作业的类型。当前包括人工标注-图像分类、人工标注-物体检测两种类型。
标注进度	当前数据标注作业的标注进度。
图片数量	当前数据标注作业的包含的图片数量。
创建时间	当前数据标注作业的创建时间。
描述	当前数据标注作业的简要描述。
来源	当前数据标注作业数据的OBS路径。
操作	对标注作业的操作。 <ul style="list-style-type: none">● “删除”：删除当前标注作业，或者删除当前标注作业及标注作业中所有的数据。● “校验”：单击“校验”，可进入校验页面。 <p>说明 当前校验功能，对数据集或图片未做任何处理。</p>

注意

数据标注即将下线，可单击页面上方“迁移”按钮将数据迁移至“数据标注-Beta”。

3.2.2 创建数据标注作业

进行数据标注之前，首先需要创建数据标注作业，具体操作步骤如下。

步骤1 单击数据标注页面左上方“创建”。

步骤2 在“创建数据标注作业”页面填写参数，填写标注作业名称、描述，选择标注类型、选择待标注作业数据的OBS路径。

说明

当前人工标注类型仅支持图像分类和物体检测两种类型的标注作业。

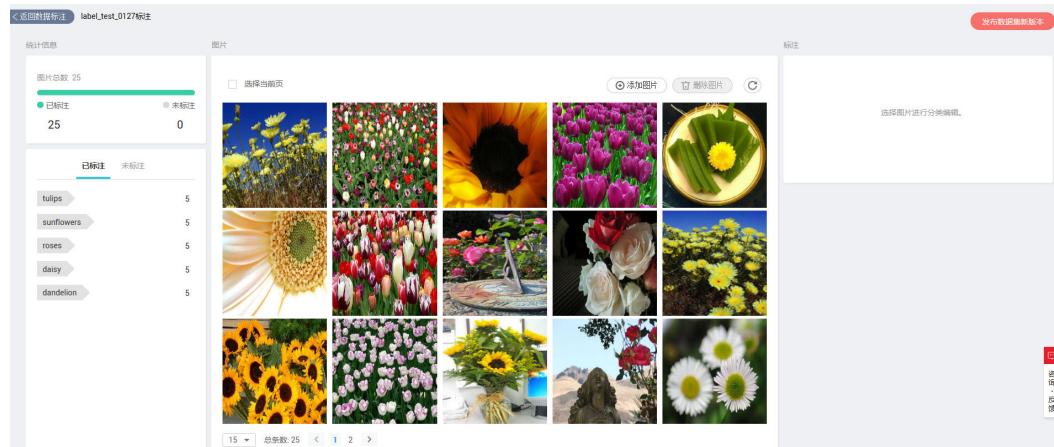
步骤3 单击“创建”，完成标注作业的创建操作。

----结束

创建数据标注作业后，单击作业名称可进入数据标注页面，图像分类标注请参见[人工标注-图像分类](#)，物体检测标注请参见[人工标注-物体检测](#)。

3.2.3 人工标注-图像分类

单击人工标注-图像分类的标注作业名称，可进入数据标注页面，如**图3-1**所示。该页面主要分为四部分内容，四个区域的内容介绍如**表3-2**所示。

图 3-1 人工标注-图像分类**表 3-2 界面内容介绍**

区域	说明
统计信息	图片的统计信息，包含图片总数、已标注图片数量及标签名称、未标注图片数量。
图片	图片的相关操作，可对图片进行添加、删除、预览操作，并可同步OBS桶中最新上传的图片。同时可以单个或批量选择图片，然后在“标注”区域完成图片标签的添加。 图片的相关详细操作请参见： <ul style="list-style-type: none">● 图片预览● 添加图片● 删除图片● 数据源同步
标注	图片标注，详细标注操作请参见 图片标注 。
发布数据集新版本	完成图片标注操作后，您可以单击页面右上角“发布数据集新版本”，系统会自动执行数据集同步操作，同步完成后页面会跳转到“数据管理>数据集”页面，您可以在“版本管理”页面中查看新发布的数据集版本。 说明 如当前数据标注作业的数据还未创建数据集，则页面右上角显示为“生成数据集”。单击“生成数据集”，系统会自动启动创建数据集任务，创建成功后页面会跳转到“数据管理>数据集”页面，创建的数据集名称与标注作业名称一致。

图片预览

单击浮于图片上方的“图片预览”，即可进行图片的预览。

图片标注

由于模型训练过程需要大量有标签的图片数据，因此在模型训练之前需对没有标签的图片添加标签。通过ModelArts您可对图片进行一键式批量添加标签，快速完成对图片的标注操作，也可以对已标注图片修改或删除标签进行重新标注，具体操作如下。

- 添加标签

- 选择未标注的图片。单击“统计信息”区域中的“未标注”，然后在“图片”区域单击浮于图片上方的“选择图片”依次选中图片，或勾选上方“选择当前页”选中该页面所有图片。
- 添加标签。选中图片后，在“标注”区域输入标签名称或从弹出的列表中选择已添加的标签，然后按Enter键添加，如图3-2所示。

图 3-2 图像分类图片标注



- 删除或修改单个图片标签

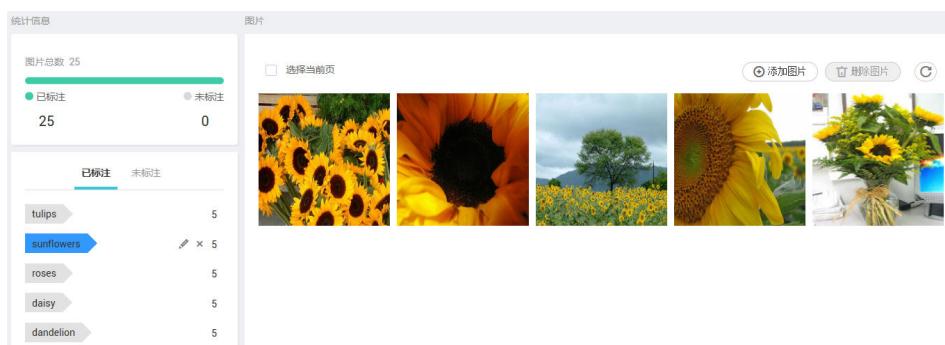
- 单击“统计信息”区域中的“已标注”，然后在“图片”区域单击浮于图片上方的“选择图片”选中图片。
- 单击标签右侧 ，删除该标签。然后输入新的标签名称，或从弹出的列表中选择已添加的标签，然后按Enter键添加，如图3-3所示。

图 3-3 删除并添加新标签



- c. 单击“确定”，完成单个图片标签的删除并添加新标签的操作。
- 批量删除或修改图片标签
 - a. 单击“统计信息”区域中的“已标注”，然后单击下方需要批量修改或删除的标签名称，如图3-4所示。

图 3-4 批量删除或修改图片标签



- b. 单击标签右侧 ，在弹出的对话框中可重命名标签名称。或者单击标签右侧 ，在弹出对话框中，可选择“仅删除标签”或“删除标签及仅包含此标签的图片”。

添加图片

通过数据添加操作，可将您本地的图片快速添加到ModelArts，同时自动上传至创建项目时所选择的OBS路径中。单击“图片”区域中的“添加”，在弹出的对话框中单击“添加文件”并选择要添加的图片，即可完成图片的添加操作。

说明

图片只支持JPG、JPEG、PNG、BMP，且一次上传所有图片的总大小不能超过8MB。

删除图片

通过数据删除操作，可将需要丢弃的图片数据快速删除。单击浮于图片上方的“选择图片”依次选中需要删除的图片，或者勾选上方“选择当前页”选中该页面所有图片，然后单击“图片”区域中“删除”，即可完成图片的删除操作。

数据源同步

为了快速获取用户OBS桶中最新图片，可单击“图片”区域右上角的同步按钮C，快速将通过OBS上传的图片数据添加到ModelArts。

3.2.4 人工标注-物体检测

单击人工标注-物体检测的标注作业名称，可进入数据标注页面，如图3-5所示。该页面主要分为三部分内容，三个区域的内容介绍如表3-3所示。

图 3-5 人工标注-物体检测

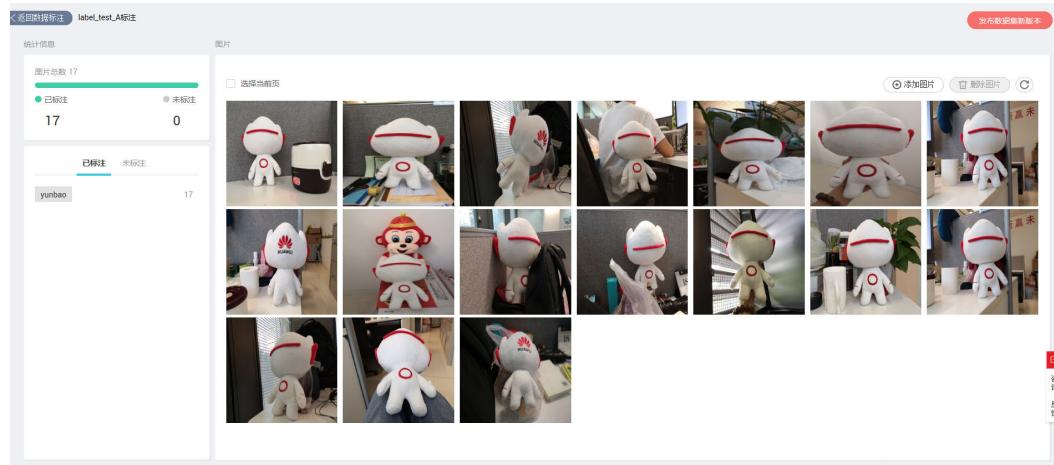


表 3-3 界面内容介绍

区域	说明
统计信息	图片的统计信息，包含图片总数、已标注图片数量及标签名称、未标注图片数量。
图片	图片的相关操作，可对图片进行添加、删除操作，并可同步OBS桶中最新上传的图片。同时可以单个或批量选择图片，进行图片的标注操作。 图片的相关详细操作请参见： <ul style="list-style-type: none">● 图片标注● 添加图片● 删除图片● 数据源同步

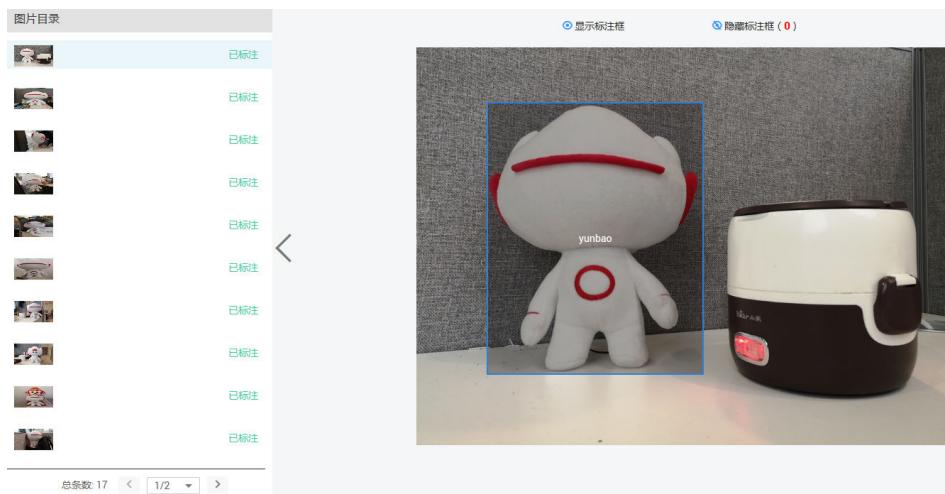
区域	说明
发布数据集新版本	<p>完成图片标注操作后，您可以单击页面右上角“发布数据集新版本”，系统会自动执行数据集同步操作，同步完成后页面会跳转到“数据管理>数据集”页面，您可以在“版本管理”页面中查看新发布的数据集版本。</p> <p>说明</p> <p>如当前数据标注作业的数据还未创建数据集，则页面右上角显示为“生成数据集”。单击“生成数据集”，系统会自动启动创建数据集任务，创建成功后页面会跳转到“数据管理>数据集”页面，创建的数据集名称与标注作业名称一致。</p>

图片标注

● 物体标注

- 单击浮于图片上方的“图片标注”，进入图片标注界面。
- 用鼠标框选图片中的物体区域，然后在弹出的对话框中输入标签名称，按Enter键添加，如图3-6所示。

图 3-6 物体检测图片标注



- 完成当前图片中所有物体标注后，可选择左侧图片目录中其他未标注图片，重复上一步骤，完成所有图片的标注。

说明

- 一张图片可添加多个标签，且仅需要标注出需要检测的物体。
- 如需要检测的物体之间相似度很高，则标注时需要将差异点置于标注框内。

● 删除或修改单个图片物体标注

- 单击“统计信息”区域中的“已标注”。
- 单击浮于图片上方的“图片标注”，进入图片标注界面。（单击“已标注”后您可以单击下方标签名称，“图片”区域将仅显示含此标签图片。）
- 单击页面右侧的 ，可删除当前物体标注，并可参见[物体标注](#)重新对图片中物体进行标注。单击页面右侧的 ，可修改当前物体标注名称。

说明

您也可以通过鼠标对已标注图片中的标注框进行调整，以调整标注框位置、大小。

- 批量修改或删除图片物体标注
 - a. 单击“已标注”，然后单击下方需要批量修改或删除的图片物体标注名称，如图3-7所示。

图 3-7 批量修改或删除图片物体标注



- b. 单击标签右侧 ，在弹出的对话框中可重命名标签名称。或者单击标签右侧 ，在弹出对话框中，可选择“仅删除标签”或“删除标签及仅包含此标签的图片”。

添加图片

通过数据添加操作，可将用户本地计算机的图片快速添加到ModelArts，同时自动上传至创建项目时所选择的OBS路径中。单击“图片”区域中的“添加”，在弹出的对话框中单击“添加文件”并选择要添加的图片，即可完成图片的添加操作。

说明

图片只支持JPG、JPEG、PNG、BMP，且一次上传所有图片的总大小不能超过8MB。

删除图片

通过数据删除操作，可将需要丢弃的图片数据快速删除。单击浮于图片上方的“选择图片”依次选中需要删除的图片，或者勾选上方“选择当前页”选中该页面所有图片，单击“图片”区域中的“删除”，即可完成图片的删除操作。

数据源同步

为了快速获取用户OBS桶中最新图片，可单击“图片”区域中的同步按钮 ，快速将通过OBS上传的图片数据添加到ModelArts。

3.3 数据集

数据管理“数据集”界面主要分为四部分内容，如图3-8所示，四个区域内容介绍如表3-4所示。

图 3-8 数据集界面

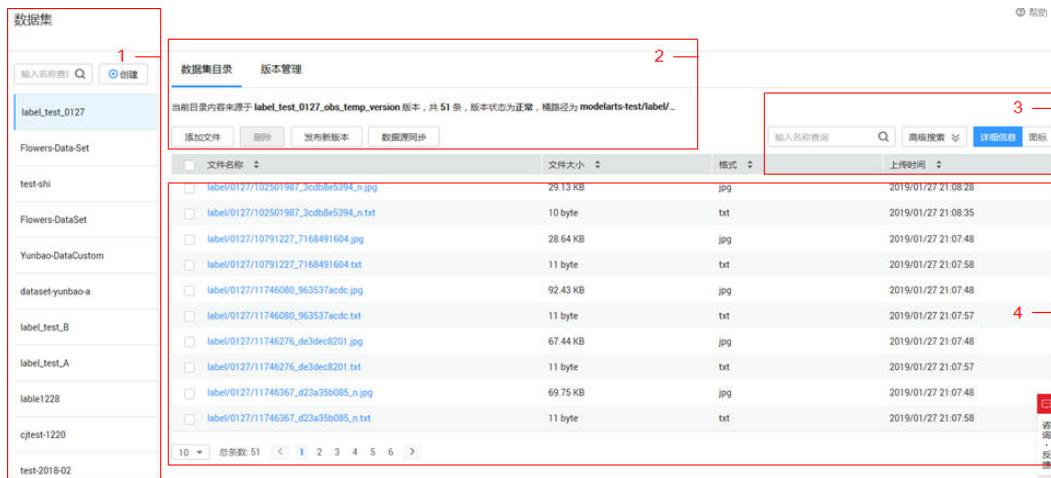


表 3-4 界面内容介绍

区域	说明
1	<p>数据集列表，列举用户所创建的数据集，同时可进行如下操作：</p> <ul style="list-style-type: none"> ● 查询：输入数据集名称单击 查询。 ● 修改：鼠标移动到相应数据集，然后单击 可修改数据集名称、描述。 ● 删除：鼠标移动到相应数据集，然后单击 可“删除当前数据集”，或者“删除数据集同时删除桶内文件”。 ● 创建：单击“创建”可创建数据集，详细步骤参见创建数据集。
2	<p>当前版本数据集信息，包括当前目录内容来源、数据数量、版本状态及桶路径。</p> <p>同时可进行如下操作：</p> <ul style="list-style-type: none"> ● 添加文件：单击“添加文件”，在弹出的对话框中选择文件，您可设置添加文件的存储位置，单击“上传文件”，完成文件的添加操作。 说明 <ul style="list-style-type: none"> ● 上传文件的总大小一次不能超过了8MB。 ● 如不设置存储位置，添加的文件将存储在当前OBS路径下。路径中任何单个斜杠 (/) 表示分隔并创建多层级的文件夹。 ● 文件夹名称不能包含“.”、“*”、“?”、“<”、“>”、“ ”，文件夹名称不能以英文句号“.”或斜杠“/”开头或结尾，文件夹的绝对路径总长度不能超过1023字符，文件夹名称不能包含两个以上相邻的斜杠“/”。 ● 删除：选中想要删除的文件，单击“删除”，完成文件的删除操作。 ● 发布新版本：单击“发布新版本”，可在弹出的对话框中填写描述，单击“确定”，完成新版本发布操作。 ● 数据源同步：单击“数据源同步”，可快速将通过OBS上传的文件数据添加到ModelArts。

区域	说明
3	可选择“详细信息”和“图标”两种显示方式。在“详细信息”显示模式下，可单击文件名称对图片文件、txt类型文件进行预览。 可输入文件名称进行简单搜索，或者输入文件大小范围、格式、上传时间段进行高级搜索。
4	当前版本数据集文件列表，包含文件名称、文件大小、格式及上传时间。

创建数据集

- 步骤1** 单击数据集页面左上角“创建”。
- 步骤2** 在弹出的对话框中输入名称、描述，并选择数据集存储路径。

说明

由于创建数据集需要开启多版本控制功能，选择数据集存储路径后，如所选择OBS桶没有开启多版本控制时，会弹出对话框提示启动[多版本控制](#)。

- 步骤3** 单击“确定”，完成数据集的创建。

----结束

版本管理

在“数据集>版本管理”页签，您可查看当前数据集版本的演进过程，如图3-9所示。版本名称自动生成，规则为“数据集名称_版本号”。数据集创建成功后，会自动生成一个临时版本，版本名称为“数据集名称_obs_temp_version”。如要切换目录，将鼠标移动到相应的版本名称上，然后单击“设置为当前目录”，即可将该版本设置为当前目录。

说明

在当前数据集目录进行的“添加文件”、“删除”操作会自动保存到临时版本中，您可以在版本管理中查看增加和删除的文件数量。

图 3-9 数据集版本管理

版本名称	状态	增加文件	删除文件	文件总数	描述	操作
label_test_B_obs_temp_version	正常	0	0	401	-	对比
label_test_B_V002	正常	0	0	401	-	对比
label_test_B_V001	正常	0	0	401	-	对比

数据集版本对比

单击版本管理页面右侧的“对比”，可进入数据集版本对比界面，如图3-10所示。

图 3-10 数据集版本对比

The screenshot shows a comparison interface for dataset versions. At the top, there are dropdown menus for '当前版本' (Current Version) set to 'bgreeff_obs_temp...' and '对比版本' (Comparison Version) set to 'bgreeff_V003'. Below this is a search bar with placeholder '输入名称查询' (Input name to search) and a '高级搜索' (Advanced Search) button.

Below the search bar is a table listing files. The columns are: 文件名称 (File Name), 文件大小 (File Size), 格式 (Format), 上传时间 (Upload Time), and 所属版本 (Belonging Version). The table contains 107 new files and 83 deleted files. A specific file row is highlighted in blue, showing its details: hard_sample/8cba6f89-f660-42af-b4a4-3223e03a49d5/images/30e376d0-73c6-4074-8c28-3322bd9485c2.jpg, 140.31KB, jpg, 2018/09/29 18:41:06, bgreeff_obs_temp_version.

文件名称	文件大小	格式	上传时间	所属版本
hard_sample/8cba6f89-f660-42af-b4a4-3223e03a49d5/images/30e376d0-73c6-4074-8c28-3322bd9485c2.jpg	140.31KB	jpg	2018/09/29 18:41:06	bgreeff_obs_temp_version
hard_sample/82d8bd8d-67cf-4cb7-bbe3-4792a1554ae7/ln...	125.20KB	jpg	2018/09/29 17:29:25	bgreeff_obs_temp_version
hard_sample/d6115519-2830-47f7-9d56-d176ac9bfee3/ln...	192.87KB	jpg	2018/09/29 18:41:07	bgreeff_obs_temp_version
hard_sample/82d8bd8d-67cf-4cb7-bbe3-4792a1554ae7/ln...	243.94KB	jpg	2018/09/29 18:56:24	bgreeff_obs_temp_version
hard_sample/3cf41491-d182-4bd6-b53f-240906ba71be/ln...	230.64KB	jpg	2018/09/29 18:56:23	bgreeff_obs_temp_version
hard_sample/82d8bd8d-67cf-4cb7-bbe3-4792a1554ae7/ln...	188.79KB	jpg	2018/09/29 18:41:07	bgreeff_obs_temp_version
hard_sample/8f707670-946f-4b19-9b40-52fedb5ede36/ima...	548.12KB	jpg	2018/09/30 09:40:40	bgreeff_obs_temp_version
hard_sample/8f707670-946f-4b19-9b40-52fedb5ede36/Ann...	1Byte	txt	2018/09/30 09:43:28	bgreeff_obs_temp_version
hard_sample/8f707670-946f-4b19-9b40-52fedb5ede36/ima...	606.34KB	jpg	2018/09/30 09:40:40	bgreeff_obs_temp_version

At the bottom left, there are navigation buttons for page numbers (10, 1, 2, 3, 4, 5, ..., 11, >).

4 开发模型

4.1 Notebook

4.1.1 Notebook 简介

ModelArts集成了基于开源的Jupyter Notebook，可为您提供在线的交互式开发调试工具。您无需关注安装配置，在ModelArts管理控制台直接使用Notebook，编写和调测模型训练代码，然后基于该代码进行模型的训练。

Jupyter Notebook是一个交互式笔记本，支持运行 40 多种编程语言。关于Jupyter Notebook的详细操作指导，请参见[Jupyter Notebook使用文档](#)。

ModelArts还提供了华为自研的分布式训练加速框架MoXing，您可以在Notebook中使用MoXing编写训练脚本，让您代码编写更加高效、代码更加简洁。MoXing使用手册请参见<https://github.com/huawei-clouds/modelarts-example/tree/master/moxing-apidoc>。

支持的 AI 引擎

每个工作环境多种AI引擎，可以在同一个Notebook实例中使用所有支持的AI引擎，不同的引擎之间可快速、方便的切换，并且有独立的运行环境。

表 4-1 AI 引擎

工作环境	AI引擎	版本
Python2	“TensorFlow”	TensorFlow-1.13.1 TensorFlow-1.8
	“MXNet”	MXNet-1.2.1
	“Caffe”	Caffe-1.0.0
	“Spark”	PySpark-2.3.2
	“Scikit-learn & XGBoost”	XGBoost-Sklearn

工作环境	AI引擎	版本
	“PyTorch”	PyTorch-1.0.0
	“Conda”	Conda-python2
Python3	“TensorFlow”	<ul style="list-style-type: none">● TensorFlow-1.13.1● TensorFlow-1.8
	“MXNet”	MXNet-1.2.1
	“Spark”	PySpark-2.3.2
	“Scikit-learn & XGBoost”	XGBoost-Sklearn
	“PyTorch”	PyTorch-1.0.0
	“Conda”	Conda-python2

使用限制

- 出于安全因素考虑，ModelArts集成的Notebook暂不开放用户root权限，可使用非特权用户jovyan或者ma-user（Multi-Engine引擎）进行操作，因此暂不能使用apt-get安装操作系统软件。
- 针对当前的AI引擎框架，Notebook仅支持单机模式训练模型。如果需要使用分布式模式训练模型，建议使用ModelArts训练作业，资源池设置多节点的方式实现。

4.1.2 创建并打开 Notebook

在开始进行模型开发前，您需要创建Notebook，并打开Notebook进行编码。

背景信息

- 创建和使用Notebook需要消耗资源，需要收费。根据您选择的资源不同，收费标准不同，针对不同类型资源的价格，详情请参见[产品价格详情](#)。
- “运行中”的Notebook将一直收费，当您不需要使用时，建议停止Notebook，避免产生不必要的费用。
- “启动中”、“停止”或“错误”状态的Notebook，无法执行打开操作。
- 一个账户最多创建10个Notebook。

创建 Notebook

- 登录ModelArts管理控制台，在左侧菜单栏中选择“开发环境>Notebook”，进入“Notebook”管理页面。
- 单击“创建”进入“创建Notebook”页面，参考[表4-2](#)填写信息。

表 4-2 参数说明

参数名称	说明
“计费方式”	按需计费。当前仅支持按需计费，无需修改。
“名称”	Notebook的名称。只能包含数字、字母、下划线和中划线，长度不能超过20位且不能为空。
“描述”	对Notebook的简要描述。
“工作环境”	<p>当前支持2种工作环境，分别为“Python2”和“Python3”，不同工作环境其对应可使用的AI引擎不同，详细支持列表请参见支持的AI引擎。</p> <p>每个工作环境多种AI引擎，可以在同一个Notebook实例中使用所有支持的AI引擎，不同的引擎之间可快速、方便的切换，并且有独立的运行环境。</p> <p>说明 ModelArts还支持Keras引擎，详细说明请参见ModelArts是否支持Keras引擎？</p>
“资源池”	可选公共资源池和专属资源池，关于ModelArts资源池的介绍和购买，请参见 资源池 。
“类型”	支持CPU和GPU两种类型。GPU性能更佳，但是相对CPU而言，费用更高。
“规格”	<p>根据选择的类型不同，可选规格也不同。</p> <ul style="list-style-type: none">● CPU规格支持：“2核8GiB”、“8核32GiB”。● GPU规格支持：“8核64GiB 1*p100”（默认），“16核128GiB 2*p100”和“32核256GiB 4*p100”规格。 <p>说明 其中，“16核128GiB 2*p100”和“32核256GiB 4*p100”规格为邀测状态，请提交工单申请使用权限。</p>

参数名称	说明
“存储配置”	<p>存储配置可选“EVS”和“OBS”。</p> <ul style="list-style-type: none">● 选择“EVS”作为存储位置 根据实际使用量设置磁盘规格。磁盘规格默认5GB。ModelArts提供5GB容量供用户免费使用。超出5GB时，超出部分每GB按“超高IO”类型的收费标准进行按需收费。磁盘规格的取值范围为5GB~500GB。 选择此模式，用户在Notebook列表的所有文件读写操作都是针对容器中的内容操作，与OBS无关；重启该实例，内容不丢失。● 选择“OBS”作为存储位置 在“存储位置”右侧单击“选择”，设置用于存储Notebook数据的OBS路径。如果想直接使用已有的文件或数据，可将数据提前上传至对应的OBS路径下。 选择此模式，用户在Notebook列表的所有文件读写操作是基于所选择的OBS路径下的内容操作，与当前实例空间无关。如果您需要将内容同步到实例空间，先选中该内容，单击“Sync OBS”，即可将所选内容同步到当前容器空间，详细操作可参见使用Sync OBS功能。重启该实例时，内容不丢失。

3. 参数填写完成后，单击“下一步”进行规格确认。
4. 参数确认无误后，单击“立即创建”，完成Notebook的创建操作。

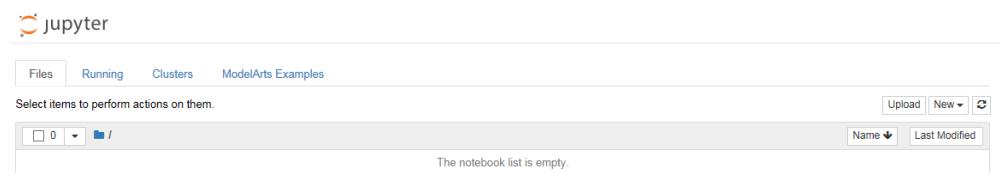
进入Notebook列表，正在创建中的Notebook状态为“启动中”，创建过程需要几分钟，请耐心等待。当Notebook状态变为“运行中”时，表示Notebook已创建完成。

打开 Notebook

在Notebook列表中，选择需要打开的Notebook，单击“操作”列中的“打开”，进入“Jupyter Notebook”开发页面。

在“Jupyter Notebook”页面中，有“Files”、“Running”、“Clusters”、“ModelArts Examples”4个页签。

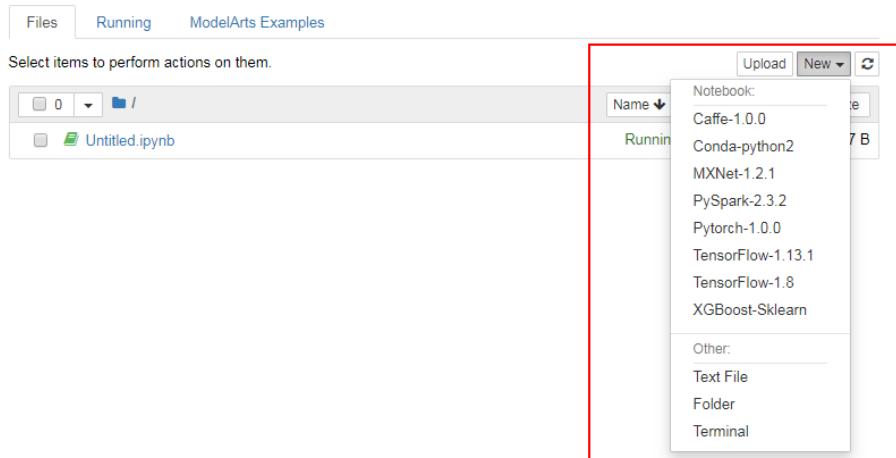
图 4-1 Jupyter Notebook 开发界面



选择不同 AI 引擎新建文件

打开Notebook实例后，进入“Jupyter Notebook”页面，在“Files”页签下，您可以单击右上角“New”，然后选择所需的AI引擎，创建一个用于编码的文件。

图 4-2 选择不同的 AI 引擎



4.1.3 访问 Notebook 并进行开发

4.1.3.1 使用 ModelArts 示例

ModelArts提供了Notebook示例，方便初学者通过示例，快速了解如何使用ModelArts Notebook。



Jupyter Notebook示例使用nbexamples扩展，有关nbexamples扩展的更多信息，请参阅<https://github.com/danielballan/nbexamples>。

预览 ModelArts Examples

1. 在Notebook列表中，创建并打开一个Notebook，或者直接打开已有的Notebook。
2. 在Jupyter页面中，单击“ModelArts Examples”页签，此页面罗列了“Machine Learning Introduction”和“ModelArts Python Sdk”的示例。ModelArts提供的所有示例说明如表4-3所示。每个示例提供了详细的说明，您可以单击右侧的“Preview”预览示例。

图 4-3 进入 ModelArts Examples

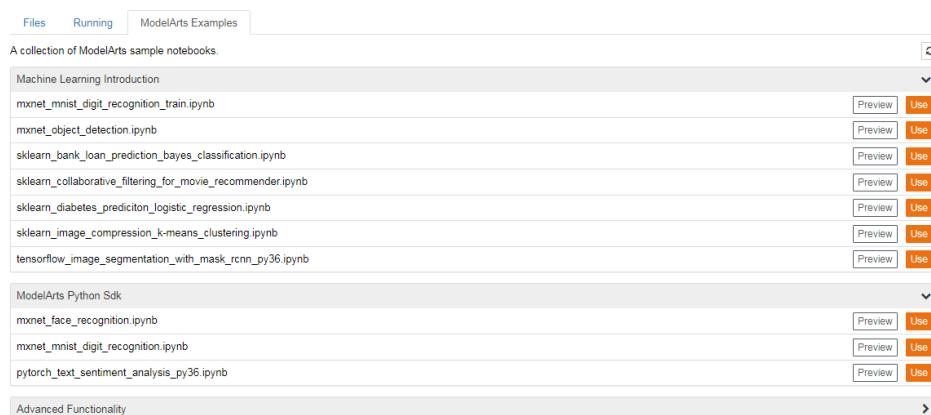


表 4-3 Examples 说明

示例名称	说明
mxnet_mnist_digit_recognition_train.ipynb	使用基于mxnet库实现一个“手写数字识别预测”功能，算法选择为反向神经网络，使用的镜像为“MXNet-1.2.1-python2.7”或者“MXNet-1.2.1-python3.6”。具体使用示例也可参见 使用Notebook实现手写数字识别 。
sklearn_bank_loan_prediction_bayes_classification.ipynb	使用基于sk-learn库实现一个“银行理财预期”的二分类，算法选择为随机森林，使用的镜像为ML-1.0.0-python27。
sklearn_diabetes_predicton_logistic_regression.ipynb	
pytorch_text_sentiment_analysis_py36.ipynb	使用基于pytorch库实现一个“文本语义分析”的功能，算法选择为循环神经网络，使用的镜像为PyTorch-1.0.0-python2.7。

使用 ModelArts Examples

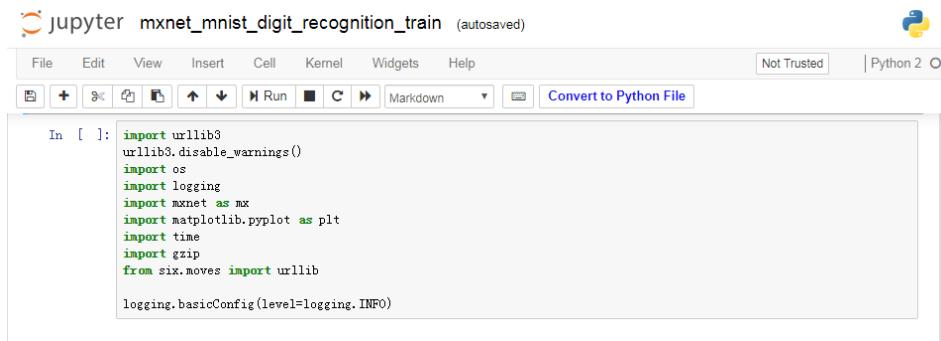
在Jupyter界面中使用或查看Notebook示例。

1. 在Notebook列表中，创建并打开一个Notebook，或者直接打开已有的Notebook。
2. 在Jupyter页面中，单击“ModelArts Examples”页签，选择需要使用的示例，单击示例右侧的“Use”。
3. 在弹出的“Create a copy in your home directory”对话框中，设置新的“ipynb”文件名称，也可以直接使用默认文件，然后单击“Create copy”保存并打开新的“ipynb”文件。打开的示例文件如图4-5所示。

使用示例是指将示例文件创建一个副本，其代码内容与示例一致。

图 4-4 创建示例副本

图 4-5 打开示例文件

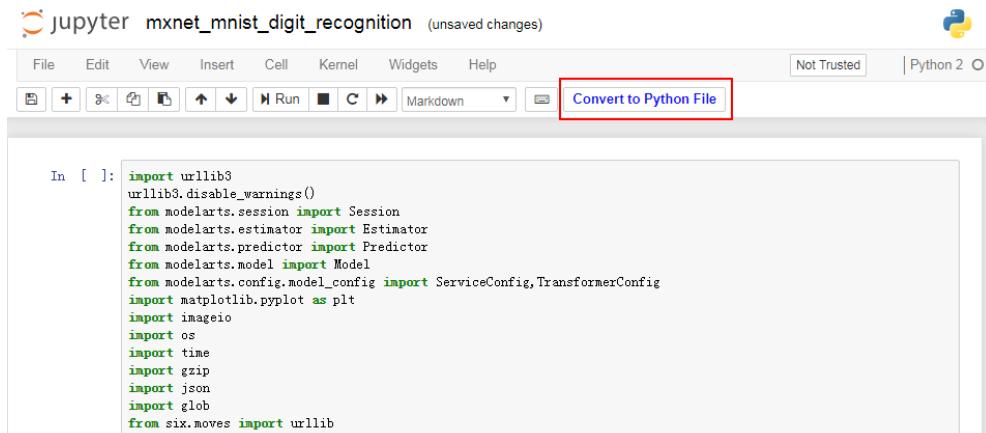


4.1.3.2 使用 Convert to Python File 功能

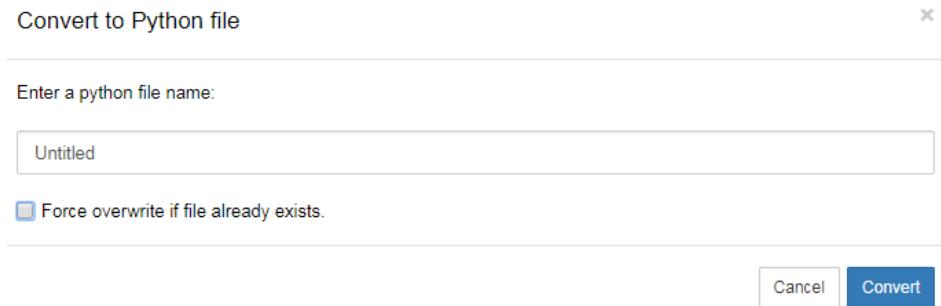
代码开发完成后，还支持将您输入的代码保存为“.py”文件，用于ModelArts训练作业的启动文件。

1. 在Notebook列表中，创建并打开一个Notebook，或者直接打开已有的Notebook。
2. 在Jupyter页面中，单击“New”，然后在列表下选择所需的AI引擎进入代码开发界面。
3. 在开发界面完成代码编写后，单击左上角的保存按钮。然后单击“Convert to Python File”将当前的“ipynb”转化为“python”文件，此功能可直接将您输入的代码保存为“.py”文件到工作目录。
生成的“.py”文件可用于ModelArts训练作业的启动文件。

图 4-6 Convert to Python File



4. 在弹出的对话框中，根据实际情况填写文件名称，然后勾选或去勾选“Force overwrite if file already exists.”，默认为不勾选，表示当目录下存在相同名称文件时，不会执行覆盖操作。然后单击“Convert”完成操作。

图 4-7 设置并保存

4.1.3.3 使用 Sync OBS 功能

如您在创建Notebook实例时，选择了OBS的“存储位置”，您写的代码会自动存储到您选择的OBS目录下。如您需要不同ipynb文件进行代码的相互调用，则可以使用Sync OBS功能。

Sync OBS功能是将在Notebook实例文件list列表选中的对象从OBS桶路径下同步到当前容器目录“~/work”下。

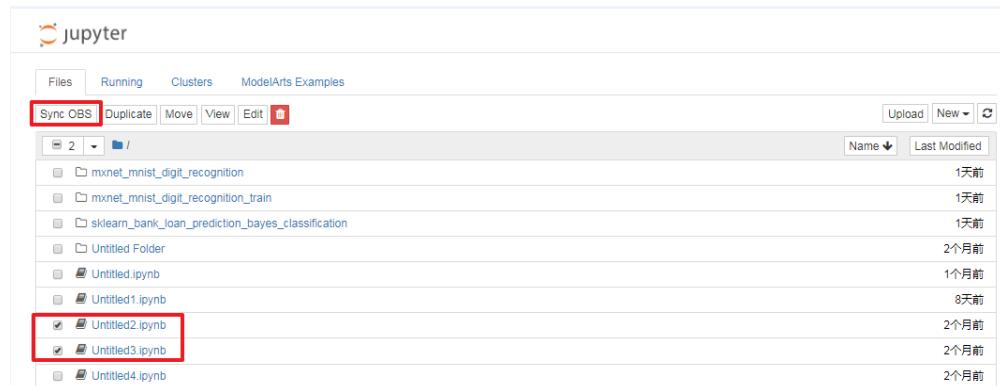
注意事项

- 单次同步文件个数最多是1024个。
- 同步的对象总大小不超过500MB，即当前容器目录“~/work”下已有200MB的文件了，那么用户使用Sync OBS能同步的最多就只有300MB文件。
- Sync OBS功能只在带有OBS存储的实例上存在，因为非OBS存储的Notebook实例，其所有的文件读写操作都在用户容器里，即在“~/work”容器目录。

操作步骤

Sync OBS功能的操作指导如下所示。

例如，“Untitled2.ipynb”需要调用“Untitled3.ipynb”中的“module”。选中这两个“ipynb”文件，然后单击界面上的“Sync OBS”，同步成功后，即可在代码间相互调用。

图 4-8 使用 Sync OBS 功能

4.1.3.4 使用 Notebook 上传大文件

在“Notebook”页面中，通过单击“Upload”按钮上传文件。当上传文件提示受限时，您可以先将大文件上传到OBS中，OBS上传文件的操作指导，请参见[上传文件](#)。然后根据不同场景将大文件下载到Notebook中。

对于挂载 EVS 的 Notebook 实例下载文件

可以使用以下任一方式将大文件下载到Notebook中：

1. 使用ModelArts SDK的OBS接口[从OBS下载数据](#)将OBS中的文件下载到Notebook后进行操作。
2. 使用[Moxing操作OBS文件](#)将OBS中的文件同步到Notebook后进行操作。

对于带 OBS 存储的 Notebook 实例下载文件

使用[使用Sync OBS功能](#)方式将OBS中的文件同步到Notebook即可。

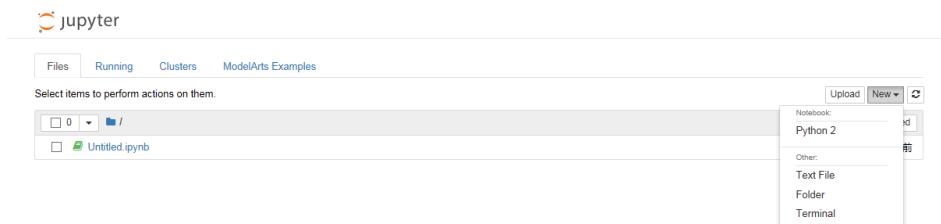
4.1.3.5 调用 ModelArts SDK

ModelArts集成的Notebook支持SDK的调用，通过调用ModelArts SDK，您可以在Notebook中轻松完成数据集管理、启动模型训练等。ModelArts的SDK详细介绍请参见《[ModelArts SDK参考](#)》。

初始化

1. 在Notebook列表中，创建并打开一个Notebook，或者直接打开已有的Notebook。
2. 在Jupyter页面中，单击“New”，然后在列表下单击“Terminal”。

图 4-9 打开 Terminal



3. 在弹出的窗口中输入`vi .modelarts/config.json`编辑配置文件，然后按“Enter”键。
4. 在打开的`config.json`配置文件中，输入相应的华为云账号与密码。然后按“Esc”键，输入“:wq”保存配置文件。至此，完成初始化操作。
 - 将红色方框内的账号信息替换成当前账号信息，如果您使用主账号登录，将“account”的参数值设置为空，如下所示：
`"account": ""`
 - 如需要在其他Notebook实例中调用SDK，需要在该实例中也完成初始化的操作。

图 4-10 设置账号及密码



打开 Python Kernel 调用 SDK

1. 进入Jupyter页面，单击“New”，然后在列表下选择对应的AI引擎创建文件。
2. 在弹出窗口中，输入SDK相关代码即可。具体SDK代码可参见《[ModelArts SDK参考](#)》。

4.1.3.6 安装外部库和内核

ModelArts Notebook中已安装Jupyter、Python程序包等多种环境，包括TensorFlow、MXNet、Caffe、PyTorch、Spark等。为了方便使用，您也可以使用pip install在Jupyter Notebook或Terminal中安装外部库。

在 Jupyter Notebook 中安装

如下操作以在Notebook中安装Shapely为例。

1. 在ModelArts管理控制台，进入“开发环境>Notebook”，并在Notebook列表中，打开一个Notebook。
2. 在Jupyter页面中，选择“New”（新建），然后选择相应的AI引擎。
3. 在代码输入栏输入以下命令安装Shapely。

!pip install Shapely

在 Terminal 中安装

如下操作以安装Shapely为例，在Notebook实例terminal中使用pip安装Shapely。

1. 在ModelArts管理控制台，进入“开发环境>Notebook”，并在Notebook列表中，打开一个Notebook。
2. 在Jupyter页面中，选择“New”（新建），然后选择“terminal”。
3. 如非Multi-Engine的实例在代码输入栏输入以下命令安装Shapely。

opt/conda/envs/python27_tf/bin/pip install Shapely

4. Multi-Engine的实例存在多个引擎，在“/home/ma-user/”路径下提供了README，可参照README切换到相应的引擎环境中安装包，如在TensorFlow-1.13.1中安装Shapely可按照如下步骤操作：

```
source /home/ma-user/anaconda3/bin/activate TensorFlow -1.13.1  
pip install Shapely
```

TensorFlow、MXNet、PyTorch、Caffe、Scikit-learn & XGBoost和Spark算法引擎在terminal中的Python路径请参见[表4-4](#)，其中pip也是在以下路径目录下。Multi-Engine的实例可参考提供的README。

表 4-4 AI 引擎及路径说明

AI引擎	版本	Python路径
TensorFlow	TF-1.8.0-python2.7	/opt/conda/envs/python27_tf/bin/python
TensorFlow	TF-1.8.0-python3.6	/opt/conda/envs/python36_tf/bin/python
MXNet	MXNet-1.2.1-pyton2.7	/opt/conda/envs/python27_mxnet/bin/python
MXNet	MXNet-1.2.1-pyton3.6	/opt/conda/envs/python36_mxnet/bin/python
PyTorch	PyTorch-1.0.0-pyton2.7	/opt/conda/envs/python27_pytorch/bin/python
PyTorch	PyTorch-1.0.0-pyton3.6	/opt/conda/envs/python36_pytorch/bin/python
Caffe	Caffe-1.0.0-pyton2.7	/opt/conda/envs/python27_caffe/bin/python
Scikit-learn & XGBoost	ML-1.0.0-pyton2.7	/opt/notebook/anaconda2/bin/python
Spark	Spark-2.2.0-pyton2.7	
Scikit-learn & XGBoost	ML-1.0.0-pyton3.6	/opt/notebook/anaconda3/bin/python
Spark	Spark-2.2.0-pyton3.6	

 **说明**

由于在创建ModelArts训练作业时，是启动一个新的独立运行环境，不会关联在Notebook环境中安装的包；因此需要在启动代码中，导入安装包前添加：os.system('pip install xxx')。

例如，在训练作业中需要使用依赖包Shapely，在Notebook中安装完成后，需要在启动代码中添加以下代码：

```
os.system('pip install Shapely')  
import Shapely
```

4.1.4 启动或停止 Notebook

由于运行中的Notebook将一直收费，您可以通过停止操作，停止收费。对于停止状态的Notebook，可通过启动操作重新使用Notebook。

登录ModelArts管理控制台，在左侧菜单栏中选择“开发环境>Notebook”，进入“Notebook”管理页面。执行如下操作停止或启动Notebook。

- **停止Notebook：**单击“操作”列的“停止”。只有处于“运行中”状态的Notebook可以执行停止操作。
- **启动Notebook：**单击“操作”列的“启动”。只有处于“停止”状态的Notebook可以执行启动操作。

4.1.5 删 除 Notebook

针对不再使用的Notebook，您可以删除Notebook以释放资源。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“开发环境>Notebook”，进入“Notebook”管理页面。
2. 在Notebook列表中，单击操作列的删除，在弹出的确认对话框中，确认信息无误，然后单击确定完成删除操作。

只有处于“停止”或“错误”状态的Notebook才可以执行删除操作。如果Notebook处于运行中，请先执行停止Notebook操作。



说明

Notebook删除后不可恢复，请谨慎操作。但是Notebook中创建的文件仍然存储在创建Notebook时指定的EVS或OBS中。

5 训练模型

5.1 模型训练简介

ModelArts提供了模型训练的功能，方便您查看训练情况并不断调整您的模型参数。您还可以基于不同的数据，选择不同规格的资源池（CPU或GPU）用于模型训练。除支持用户自己开发的模型外，ModelArts还提供了预置算法，您可以不关注模型开发，直接使用预置算法，通过算法参数的调整，得到一个满意的模型。

模型训练功能说明

表 5-1 功能说明

功能	说明	详细指导
预置算法	ModelArts基于业界常用的AI引擎，提供了常见用途的算法，并预置在系统中。您可以不关注模型开发，直接选择此算法用于训练作业即可。	预置算法简介
训练作业管理	支持创建训练作业、查看训练作业详情、管理训练作业版本、查看模型训练溯源图并且支持查看评估详情。	创建训练作业 管理训练作业版本 查看作业详情
作业参数管理	您可以将某一个训练作业的参数配置保存为作业参数，包含数据来源、算法来源、运行参数、资源池等参数信息，已保存的作业参数，可一键式应用到创建新的训练作业，大大提高效率。	管理作业参数
模型训练可视化（TensorBoard）	TensorBoard是一个可视化工具，能够有效地展示TensorFlow在运行过程中的计算图、各种指标随着时间的变化趋势以及训练中使用到的数据信息。TensorBoard当前只支持基于TensorFlow和MXNet引擎的训练作业。	管理TensorBoard

5.2 预置算法简介

ModelArts基于业界常用的AI引擎，提供了常见用途的算法，并预置在系统中。您可以不关注模型开发，直接选择此算法用于训练作业即可。

目前，ModelArts提供的预置算法主要是基于MXNet和TensorFlow两种引擎，其用途主要覆盖检测物体类别和位置、图像分类、图像语义分割的场景。

查看预置算法

在ModelArts管理控制台，选择“训练作业”，然后单击“预置算法”页签，进入预置算法列表页面。在预置算法列表中，您可以单击算法名称前的▼，查看该算法的详情。

另外，您还可以在算法的“操作”列，单击“创建训练”，快速创建一个训练作业，将此算法用于训练作业中的“算法来源”。



说明

使用预置算法创建训练前，需准备好训练数据并上传至OBS，数据存储路径及数据格式要求请参见[管理数据（推荐使用）](#)。

图 5-1 预置算法列表

名称	用途	引擎类型	精度	大小 (MB)	操作
yolo_v3	检测物体类别和位置	MXNet, MXNet-1.2.1-pyth...	81.7%(mAP)	235.31	创建训练
retinanet_resnet_v1_50	检测物体类别和位置	TensorFlow, TF-1.8.0-pyth...	83.15%(mAP)	255.15	创建训练
inception_v3	图像分类	TensorFlow, TF-1.8.0-pyth...	78.00%(top1), 93.90%(t...	103.78	创建训练
darknet_53	图像分类	MXNet, MXNet-1.2.1-pyth...	78.56%(top1), 94.43%(t...	158.89	创建训练
SegNet_VGG_BN_16	图像语义分割	MXNet, MXNet-1.2.1-pyth...	89%(pixel acc)	112.41	创建训练
ResNet_v2_50	图像分类	MXNet, MXNet-1.2.1-pyth...	75.55%(top1), 92.6%(to...	97.76	创建训练
ResNet_v1_50	图像分类	TensorFlow, TF-1.8.0-pyth...	74.2%(top1), 91.7%(top5)	200.84	创建训练
Faster_RCNN_ResNet...	检测物体类别和位置	MXNet, MXNet-1.2.1-pyth...	80.05%(mAP)	182.09	创建训练
Faster_RCNN_ResNet...	检测物体类别和位置	TensorFlow, TF-1.8.0-pyth...	73.6%(mAP)	281.16	创建训练

预置算法的详细介绍请参见：

- [yolo_v3](#)
- [retinanet_resnet_v1_50](#)
- [inception_v3](#)
- [darknet_53](#)
- [SegNet_VGG_BN_16](#)
- [ResNet_v2_50](#)
- [ResNet_v1_50](#)
- [Faster_RCNN_ResNet_v2_50](#)
- [Faster_RCNN_ResNet_v1_50](#)

yolo_v3

表 5-2 算法介绍

参数	说明
名称	yolo_v3
用途	检测物体类别和位置
引擎类型	MXNet, MXNet-1.2.1-python2.7
精度	81.7%(mAP) mAP是物体检测算法中衡量算法效果的指标。对于物体检测任务，每一类object都可以计算出其精确率（Precision）和召回率（Recall），在不同阈值下多次计算/试验，每个类都可以得到一条P-R曲线，曲线下的面积就是average。
训练数据集	Pascal VOC2007, 20类物体检测
数据格式	shape: [H>=224, W>=224, C>=1]; type: int8
运行参数	lr=0.0001 ; mom=0.9 ; wd=0.0005 更多可使用的运行参数，请参见 表5-3 。

表 5-3 运行参数说明

可选参数	参数说明	默认值
lr	更新参数的学习率。	0.0001
mom	训练网络的动量参数。	0.9
wd	参数权重衰减系数，L2。	0.0005
num_classes	训练中使用的图片总的类别数，这里不需要+1。	无
split_spec	训练集和验证集切分比例。	0.8
batch_size	每次更新训练的图片数量（所有GPU总和）。	4
eval_frequency	对模型做验证的频率，默认为每个epoch都做。	1
num_epoch	训练的epoch数。	10
num_examples	参与训练的图片总数，比如图片总数是1000，由于切分数据集这里是800。	16551
disp_batches	每N个batch显示一次计算模型的loss和训练速度。	20
warm_up_epochs	warm_up策略达到指定lr的epoch数。	0

可选参数	参数说明	默认值
lr_steps	multifactor策略中lr衰减的epoch数， 默认认为在第10和15个epoch衰减为原来的0.1倍。	10,15

retinanet_resnet_v1_50

表 5-4 算法介绍

参数	说明
名称	retinanet_resnet_v1_50
用途	检测物体类别和位置
引擎类型	TensorFlow, TF-1.8.0-python2.7
精度	83.15%(mAP) mAP是物体检测算法中衡量算法效果的指标。对于物体检测任务，每一类object都可以计算出其精确率（Precision）和召回率（Recall），在不同阈值下多次计算/试验，每个类都可以得到一条P-R曲线，曲线下的面积就是average。
训练数据集	ImageNet-1k; [H, W, C=3]
数据格式	shape: [H, W, C>=1]; type: int8
运行参数	默认算法未设置运行参数，更多可使用的运行参数，请参见 表5-5 。

表 5-5 运行参数说明

可选参数	参数说明	默认值
split_spec	训练集和验证集切分比例。	train:0.8,eval:0.2
num_gpus	使用的GPU个数。	1
batch_size	每次迭代训练的图片数量（单卡）。	32
eval_batch_size	验证时每步读取的图片数量（单卡）。	32
learning_rate_strategy	训练的学习率策略（“10:0.001,20:0.0001”代表0-10个epoch学习率0.001， 10-20epoch学习率0.0001）。	0.002
evaluate_every_n_epochs	每训练N个epoch做一次验证。	1
save_interval_secs	保存模型的频率（单位：s）。	2000000

可选参数	参数说明	默认值
max_epochs	最大训练的epoch数。	100
log_every_n_steps	每训练N步打印一次日志， 默认为10步打印一次。	10
save_summaries_steps	每5步保存一次summary。	5
weight_decay	模型权重衰减的L2系数。	0.00004
optimizer	优化器。可选值为： ● dymomentumw ● sgd ● adam ● momentum	momentum
momentum	优化器参数momentum。	0.9
patience	8个epoch之后若精度相较之前的没有提升学习率会衰减。	8
decay_patience	学习率下降N次后训练停止。	1
decay_min_delta	不同学习率的指标之间的最小差值， 大于0.001， 代表精度有提升。	0.001
min_delta	监测指标若小于0.001， 代表训练无提升。	0.001
rcnn_iou_threshold	ssd和faster rcnn计算map时候使用的IOU阈值。	0.5

inception_v3

表 5-6 算法介绍

参数	说明
名称	inception_v3
用途	图像分类
引擎类型	TensorFlow, TF-1.8.0-python2.7
精度	78.00%(top1), 93.90%(top5) ● top1是指对于一个图片，如果概率最大的是正确答案，才认为正确。 ● top5是指对于一个图片，如果概率前五中包含正确答案，即认为正确。
训练数据集	imagenet, 1000类图像分类

参数	说明
数据格式	shape: [H, W, C>=1]; type: int8
运行参数	batch_size=32 ; split_spec=train:0.8,eval:0.2 ; 更多可使用的运行参数请参见 表5-7 。

表 5-7 运行参数说明

可选参数	参数说明	默认值
split_spec	训练集和验证集切分比例。	train:0.8,eval:0.2
num_gpus	使用的GPU个数。	1
batch_size	每次迭代训练的图片数量（单卡）。	32
eval_batch_size	验证时每步读取的图片数量（单卡）。	32
learning_rate_strategy	训练的学习率策略（“10:0.001,20:0.0001”代表0-10个epoch学习率0.001，10-20epoch学习率0.0001）。	0.002
evaluate_every_n_epochs	每训练N个epoch做一次验证。	1
save_interval_secs	保存模型的频率（单位：s）。	2000000
max_epoches	最大训练的epoch数。	100
log_every_n_steps	每训练N步打印一次日志，默认为10步打印一次。	10
save_summaries_steps	每5步保存一次summary。	5
weight_decay	模型权重衰减的L2系数。	0.00004
optimizer	优化器。可选值： ● dymomentumw ● sgd ● adam ● momentum	momentum
momentum	优化器参数momentum。	0.9
patience	8个epoch之后若精度相较之前的没有提升学习率会衰减。	8
decay_patience	学习率下降N次后训练停止。	1
decay_min_delta	不同学习率的指标之间的最小差值，大于0.001，代表精度有提升。	0.001

可选参数	参数说明	默认值
min_delta	监测指标若小于0.001，代表训练无提升。	0.001
image_size	训练传入模型的图片大小，为None时获取模型的默认图片大小。	None
lr_warmup_strategy	warmup策略（线性或指数）。	linear
num_readers	数据读取的线程数。	64
fp16	是否使用fp16精度训练。	FALSE
max_lr	最大学习率（dymomentum, dymomentumw 优化器以及使用use_lr_schedule时调节的学习率最大值）。	6.4
min_lr	最小学习率（dymomentum, dymomentumw 优化器以及使用use_lr_schedule时调节的学习率最小值）。	0.005
warmup	warmup在总训练步数中的占比（use_lr_schedule为lcd、poly时使用）。	0.1
cooldown	在warmup中学习率能下降到的最小值。	0.05
max_mom	最大momentum（动态momentum使用）。	0.98
min_mom	最小momentum（动态momentum使用）。	0.85
use_lars	是否使用Lars。	FALSE
use_nesterov	是否使用Nesterov Momentum。	TRUE
preprocess_threads	图片预处理的线程数。	12
use_lr_schedule	学习率调整策略 ('lcd':linear_cosine_decay, 'poly':polynomial_decay)。	None

darknet_53

表 5-8 算法介绍

参数	说明
名称	darknet_53
用途	图像分类
引擎类型	MXNet, MXNet-1.2.1-python2.7

参数	说明
精度	78.56%(top1), 94.43%(top5) ● top1是指对于一个图片，如果概率最大的是正确答案，才认为正确。 ● top5是指对于一个图片，如果概率前五中包含正确答案，即认为正确。
训练数据集	imagenet, 1000类图像分类
数据格式	shape: [H>=224, W>=224, C>=1]; type: int8
运行参数	split_spec=0.8 ; batch_size=4 ; 更多可使用的运行参数，请参见 表5-9 。

表 5-9 运行参数说明

可选参数	参数说明	默认值
num_classes	训练中使用的图片总的类别数。	无
num_epoch	训练的epoch数。	10
batch_size	每次更新参数输入的数据数量（总）。	4
lr	更新参数的学习率。	0.0001
image_shape	输入模型图片的shape。	3,224,224
split_spec	训练集和验证集切分比例。	0.8
save_frequency	保存模型的频率，即隔N个epoch保存一次模型。	1

SegNet_VGG_BN_16

表 5-10 算法介绍

参数	说明
名称	SegNet_VGG_BN_16
用途	图像语义分割
引擎类型	MXNet, MXNet-1.2.1-python2.7
精度	89%(pixel acc) pixel acc是指标记正确的像素占总像素的比例。
训练数据集	Camvid
数据格式	shape: [H=360, W=480, C==3]; type: int8

参数	说明
运行参数	deploy_on_terminal=False; deploy_on_terminal=False 更多可使用的运行参数, 请参见 表5-11 。

表 5-11 运行参数说明

可选参数	参数说明	默认值
lr	更新参数的学习率。	0.0001
mom	训练网络的动量参数。	0.9
wd	权重衰减系数。	0.0005
num_classes	训练中使用的图片总的类别数, 这里不需要+1。	11
batch_size	每次更新训练的图片数量(所有GPU总和)。	8
num_epoch	训练的epoch数。	15
save_frequency	保存模型的频率, 即隔N个epoch保存一次模型。	1
num_examples	参与训练的图片总数train.txt中的文件数。	2953
data_shape	输入模型图片的shape。	3,256,256
optimizer	优化器, 默认为随机梯度下降, 可选nag。	sgd
lr_steps	multifactor策略中lr衰减的epoch数, 默认认为在第10和15个epoch衰减为原来的0.1倍。	7,12

ResNet_v2_50

表 5-12 算法介绍

参数	说明
名称	ResNet_v2_50
用途	图像分类
引擎类型	MXNet, MXNet-1.2.1-python2.7
精度	75.55%(top1), 92.6%(top5) ● top1是指对于一个图片, 如果概率最大的是正确答案, 才认为正确。 ● top5是指对于一个图片, 如果概率前五中包含正确答案, 即认为正确。

参数	说明
训练数据集	imagenet, 1000类图像分类
数据格式	shape: [H>=32, W>=32, C>=1]; type: int8
运行参数	split_spec=0.8 ; batch_size=4 ; 更多可使用的运行参数与“darknet_53”算法一致，详情请参见 表5-9 。

ResNet_v1_50

表 5-13 算法介绍

参数	说明
名称	ResNet_v1_50
用途	图像分类
引擎类型	TensorFlow, TF-1.8.0-python2.7
精度	74.2%(top1), 91.7%(top5) ● top1是指对于一个图片，如果概率最大的是正确答案，才认为正确。 ● top5是指对于一个图片，如果概率前五中包含正确答案，即认为正确。
训练数据集	imagenet, 1000类图像分类
数据格式	shape: [H>=600,W<=1024,C>=1];type:int8
运行参数	batch_size=32 ; split_spec=train:0.8,eval:0.2 ; 更多可使用的运行参数与“inception_v3”算法一致，详情请参见 表5-7 。

Faster_RCNN_ResNet_v2_50

表 5-14 算法介绍

参数	说明
名称	Faster_RCNN_ResNet_v2_50
用途	检测物体类别和位置
引擎类型	MXNet, MXNet-1.2.1-python2.7

参数	说明
精度	80.05%(mAP) mAP是物体检测算法中衡量算法效果的指标。对于物体检测任务，每一类object都可以计算出其精确率（Precision）和召回率（Recall），在不同阈值下多次计算/试验，每个类都可以得到一条P-R曲线，曲线下的面积就是average。
训练数据集	Pascal VOC2007, 20类物体检测
数据格式	shape: [H, W, C==3]; type: int8
运行参数	lr=0.0001 ; eval_frequency=1 ; 更多可使用的运行参数，请参见 表5-15 。

表 5-15 运行参数说明

可选参数	参数说明	默认值
num_classes	训练中使用的图片总的类别数，这里需要+1，因为有一个额外的背景类。	无
eval_frequency	对模型做验证的频率， 默认为每个epoch都做。	1
lr	更新参数的学习率。	0.0001
mom	训练网络的动量参数。	0.9
wd	参数权重衰减系数， L2。	0.0005
export_model	是否将生成的模型生成为部署推理服务需要的格式。	TRUE
split_spec	训练集和验证集切分比例。	0.8
optimizer	优化器， 默认为随机梯度下降， 可选nag。	sgd

Faster_RCNN_ResNet_v1_50

表 5-16 算法介绍

参数	说明
名称	Faster_RCNN_ResNet_v1_50
用途	检测物体类别和位置
引擎类型	TensorFlow, TF-1.8.0-python2.7

参数	说明
精度	73.6%(mAP) mAP是物体检测算法中衡量算法效果的指标。对于物体检测任务，每一类object都可以计算出其精确率（Precision）和召回率（Recall），在不同阈值下多次计算/试验，每个类都可以得到一条P-R曲线，曲线下的面积就是average。
训练数据集	Pascal VOC2007, 20类物体检测
数据格式	shape: [H>=600,W<=1024,C>=1];type:int8
运行参数	batch_size=32 ; split_spec=train:0.8,eval:0.2 ; 更多可使用的运行参数与“retinanet_resnet_v1_50”算法一致，详情请参见 表5-5 。

5.3 创建训练作业

数据准备完成后，您可以创建一个训练作业，对已有数据进行模型训练。每一个训练作业创建完成后，将自动完成一次训练。

前提条件

- 数据已完成准备：已在ModelArts中创建可用的数据集，或者您已将用于训练的数据集上传至OBS目录。
- 如果“算法来源”为“常用框架”，请准备好训练脚本，并上传至OBS目录。
- 如果“算法来源”为“自定义”，请按照规范完成镜像制作，并上传至SWR服务，同时，训练脚本已上传至OBS目录。
- 已在OBS创建至少1个空的文件夹，用于存储训练输出的内容。
- 由于训练作业运行需消耗资源，确保账户未欠费。

注意事项

训练作业指定的数据集目录中，用于训练的数据名称（如图片名称、音频文件名、标注文件名称等），名称长度限制为0~255英文字符。如果数据集目录下，部分数据的文件名称超过255英文字符，训练作业将不会使用此数据，使用符合要求的数据进行继续进行训练。如果数据集目录下，所有数据的文件名称都超过了255英文字符，导致训练作业无数据可用，则会最终导致训练作业失败。

创建训练作业

1. 登录ModelArts管理控制台，在左侧导航栏中选择“训练作业”，默认进入“训练作业”列表。
2. 在训练作业列表中，单击左上角“创建”，进入“创建训练作业”页面。
3. 在创建训练作业页面，填写训练作业相关参数，然后单击“下一步”。
 - a. 填写基本信息。基本信息包含“计费模式”、“名称”、“版本”和“描述”。其中“计费模式”当前仅支持“按需计费”，不支持修改。“版本”信息由系统自动生成，按“V001”、“V002”规则命名，用户无法修改。
您可以根据实际情况填写“名称”和“描述”信息。

图 5-2 训练作业基本信息

* 计费模式 按需计费

* 名称 trainjob-test

版本 V0001 版本信息为自动生成

描述

0/256

b. 填写作业参数。包含数据来源、算法来源等关键信息，详情请参见表5-17。

图 5-3 设置作业参数

一键式参数配置 如您已保存过参数配置，可单击 [这里](#) 快速导入已保存的作业参数配置。

* 数据来源 ② 数据集 /test-modelarts2/train-mnist/ 选择 ✎

* 算法来源 预置算法 常用框架 自定义 ⚙️MoXing手册

选择常用引擎创建训练作业。

* AI引擎 TensorFlow TF-1.8.0-python2.7

* 代码目录 ② /test-modelarts2/train-mnist/ 选择

* 启动文件 ② /test-modelarts2/train-mnist/train_mnist.py 选择

运行参数 ② + 增加运行参数

* 训练输出位置 ② /test-modelarts2/mnist-model/ 选择

请尽量选择空目录来作为训练输出路径。

作业日志路径 ② /test-modelarts2/train-log/ 选择 清除

日志默认保存在服务，会不定期清除，请选择相应路径用来上传日志。

表 5-17 作业参数说明

参数名称	子参数	说明
一键式参数配置	-	如果您在ModelArts已保存作业参数，您可以根据界面提示，选择已有的作业参数，快速完成训练作业的参数配置。

参数名称	子参数	说明
数据来源	数据集	<p>从ModelArts数据管理中选择可用的数据集及其版本。</p> <ul style="list-style-type: none">● “选择数据集”：从右侧下拉框中选择ModelArts系统中已有的数据集。当ModelArts无可用数据集时，此下拉框为空。● “选择版本”：根据“选择数据集”指定的数据集选择其版本。 <p>一个训练作业，支持选择多个数据集，单击增加一个数据集，单击删除当前行指定的数据集。</p>
数据来源	数据存储位置	<p>从OBS桶中选择训练数据。在“数据存储位置”右侧，单击“选择”，从弹出的对话框中，选择数据存储的OBS桶及其文件夹。</p> <p>当“算法来源”选择“常用框架”时，一个训练作业，支持选择多个数据存储路径，单击增加一个数据存储路径，单击删除当前行指定的数据存储路径。</p>
算法来源	预置算法	使用ModelArts的预置算法，详细介绍请参见 预置算法简介 。

参数名称	子参数	说明
	常用框架	<p>选择“AI引擎”和“版本”，选择“代码目录”及“启动文件”。选择的AI引擎和编写训练代码时选择的框架必须一致。例如编写训练代码使用的是TensorFlow，则在创建训练作业时也要选择TensorFlow。</p> <p>当前ModelArts支持的AI引擎及对应版本如下所示。</p> <ul style="list-style-type: none">● TensorFlow: TF-1.8.0-python3.6、TF-1.8.0-python2.7、TF-1.13.1-python3.6、TF-1.13.1-python2.7● MXNet: MXNet-1.2.1-python3.6、MXNet-1.2.1-python2.7● Caffe: Caffe-1.0.0-python2.7● Spark_MLlib: Spark-2.3.2-python2.7、Spark-2.3.2-python3.6● Scikit_Learn: Scikit_Learn-0.18.1-python2.7、Scikit_Learn-0.18.1-python3.6● XGBoost: XGBoost-0.8-python2.7、XGBoost-0.8-python3.6● PyTorch: PyTorch-1.0.0-python2.7、PyTorch-1.0.0-python3.6 <p>说明</p> <p>MoXing是华为云ModelArts团队自研的分布式训练加速框架，它构建于开源的深度学习引擎TensorFlow、MXNet、PyTorch、Keras之上，详细说明请参见MoXing 使用说明。如果您使用的是MoXing框架编写训练脚本，在创建训练作业时，请根据您选用的接口选择其对应的AI引擎和版本。“efficient_ai”是华为云ModelArts团队自研的加速压缩工具，它支持对训练作业进行量化、剪枝和蒸馏来加速模型推理速度，详细说明请参见efficient_ai使用说明。</p>
	自定义	可使用自定义镜像创建训练作业，如何制作自定义镜像请参见 构建自定义镜像 。
运行参数	-	代码中的命令行参数设置值，请确保参数名称和代码的参数名称保持一致。 例如：train_steps=10000，其中“train_steps”为代码中的某个传参。
训练输出位置	-	选择训练结果的存储位置。 说明 为避免出现错误，建议选择一个空目录用作“训练输出位置”。请勿将数据集存储的目录作为训练输出位置。
作业日志路径	-	选择作业运行中产生的日志文件存储路径。

c. 选择用于训练作业的资源。

图 5-4 选择资源



表 5-18 资源参数说明

参数名称	说明
资源池	<p>选择训练作业资源池。训练作业支持选择公共资源池和专属资源池。</p> <p>公共资源池又可以选择CPU或GPU两种规格，不同规格的资源池，其收费标准不同，详情请参见价格详情说明。专属资源池的创建请参见资源池。</p> <p>说明</p> <p>如果您在训练代码使用的是GPU资源，则在选择资源池时只能选择GPU集群，否则会导致训练作业失败。</p>
计算节点个数	<p>选择计算节点的个数。如果节点个数设置为1，表示后台的计算模式是单机模式；如果节点个数设置大于1，表示后台的计算模式为分布式的。请根据实际编码情况选择计算模式。</p> <p>当“常用框架”选择Caffe时，只支持单机模式，即“计算节点个数”必须设置为“1”。针对其他“常用框架”，您可以根据业务情况选择单机模式或分布式模式。</p>

- d. 配置订阅消息，并设置是否将当前训练作业中的参数保存为作业参数。

图 5-5 配置订阅消息

表 5-19 订阅消息及作业参数参数说明

参数名称	说明
订阅消息	<p>订阅消息使用消息通知服务，在事件列表中选择需要监控的资源池状态，在事件发生时发送消息通知。</p> <p>此参数为可选参数，您可以根据实际情况设置是否打开开关。如果开启订阅消息，请根据实际情况填写如下参数。</p> <ul style="list-style-type: none">● “主题名”：订阅消息主题名称。您可以单击创建主题，在消息通知服务中创建主题。● “事件列表”：订阅事件。当前可选择“OnJobRunning”、“OnJobSucceeded”、“OnJobFailed”三种事件，分别代表训练运行中、运行成功、运行失败。
保存作业参数	<p>勾选此参数，表示将当前训练作业设置的作业参数保存，方便后续一键复制使用。</p> <p>勾选“保存训练参数”，然后填写“作业参数名称”和“作业参数描述”，即可完成当前参数配置的保存。训练作业创建成功后，您可以从ModelArts的作业参数列表中查看保存的信息，详细操作指导请参见管理作业参数。</p>

- e. 完成参数填写后，单击“下一步”。
4. 在“规格确认”页面，确认填写信息无误后，单击“立即创建”，完成训练作业的创建。训练作业一般需要运行一段时间，根据您选择的数据量和资源不同，训练时间将耗时几分钟到几十分钟不等。

说明

训练作业创建完成后，将立即启动，运行过程中将按照您选择的资源按需计费。
您可以前往训练作业列表，查看训练作业的基本情况。在训练作业列表中，刚创建的训练作业“状态”为“初始化”，当训练作业的“状态”变为“运行成功”时，表示训练作业运行结束，其生成的模型将存储至对应的“训练输出位置”中。当训练作业的“状态”变为“运行失败”时，您可以单击训练作业的名称，进入详情页面，通过查看日志等手段处理问题。

停止训练作业

在训练作业列表中，针对“运行中”的训练作业，您可以单击“操作”列的“停止”，停止正在运行中的训练作业。

训练作业停止后，ModelArts将停止计费。如果停止的训练作业已勾选“保存作业参数”，其设置的作业参数将继续保存至“作业参数管理”页面中。

运行结束的训练作业，如“运行成功”、“运行失败”的作业，不涉及“停止”操作。只有“运行中”的训练作业支持“停止”操作。

删除训练作业

当已有的训练作业不再使用时，您可以删除训练作业。

处于“运行中”、“运行成功”、“运行失败”、“已取消”、“部署中”状态的训练作业，您可以单击“操作”列的“删除”，删除对应的训练作业。

如果删除的训练作业已勾选“保存作业参数”，其设置的作业参数将继续保存至“作业参数管理”页面中。

5.4 管理训练作业版本

在模型构建过程中，您可能需要根据训练结果，不停的调整数据、训练参数或模型，以获得一个满意的模型。因此，ModelArts为了方便用户在调整内容后快速高效的训练模型，提供了管理训练作业版本的能力。每训练一次，生成一个版本，不同的作业版本之间，能快速进行对比，获得对比结果。

查看训练作业版本

1. 登录ModelArts管理控制台，在左侧导航栏中选择“训练作业”，默认进入“训练作业”列表。
2. 在训练作业列表中，单击训练作业名称，进入训练作业的详情页面。

默认打开最近一个版本的基本信息。当版本较多时，您可以单击左上角  **版本过滤** 过滤某几个版本进行查看。单击版本左侧的  打开作业的详细信息。训练作业的详细信息说明请参见[训练作业详情](#)。

图 5-6 查看训练作业版本



配置信息	日志	资源占用情况	评估结果
名称	trainjob-mnist-tf jobb2f5aa09	AI引擎	TensorFlow TF-1.8.0-python2.7
状态	运行成功	代码目录	/test-modelarts2/tf-mnist/codes/
运行版本	V0002	启动文件	/test-modelarts2/tf-mnist/codes/train_mnist_tf.py
开始运行时间	2019/05/29 16:52:06	训练数据集	dataset-mnist(old) dataset-mnist_V001
运行时间	00:00:50	主要运行参数	-
资源池	Computing GPU(P100) instance	训练输出位置	/test-modelarts2/tf-mnist/output/V0002/
计算节点个数	1	描述	-
日志输出位置	-	NAS 地址	-
NAS 挂载路径	-		

版本对比

在“版本管理”页面中，针对当前训练作业的所有版本，或者使用过滤功能筛选后的版本，单击右侧“查看对比结果”，可查看训练版本之间的对比，包含“运行参数”、“F1值”、“召回率”、“精确率”、“准确率”。

图 5-7 训练版本对比



版本	运行参数	F1值	召回率	精确率	准确率
V0002					
V0001					

基于训练作业版本的快捷操作

在训练作业的版本管理页面，ModelArts提供了一些快捷操作的入口，方便您在模型训练结束后，快速进行下一步操作。

表 5-20 快捷操作说明

操作	说明
创建TensorBoard	基于当前训练版本创建TensorBoard，详细参见 管理TensorBoard 。 说明 TensorBoard目前只支持TensorFlow和MXNet引擎，只有使用TensorFlow或MXNet引擎的训练作业才可以创建TensorBoard作业。
创建模型	基于当前训练版本创建模型，详细参见 导入模型 。只有“运行成功”的训练作业，支持此操作。
修改训练作业	如果当前版本的训练结果不满足业务需求时，或者训练作业“运行失败”时，您可以单击“修改”，跳转至训练作业参数设置页面，训练作业的参数说明请参见 创建训练作业 。根据实际情况调整作业参数后，单击“确定”启动新版本的训练作业。
保存作业参数	将此版本的作业参数可保存为新的作业参数。单击“更多操作>保存作业参数”，进入“作业参数”页面，确认信息无误后的，单击确定完成操作。作业参数管理详情请参见 管理作业参数 。
停止	单击“更多操作>停止”可停止当前版本的训练作业。只有“运行中”的训练作业版本才支持停止操作。
删除	单击“更多操作>删除”可停止当前版本的训练作业。

图 5-8 快捷操作



5.5 查看作业详情

训练作业运行结束后，除了管理训练作业版本之外，您可以通过查看[训练作业详情](#)、查看[溯源图](#)、查看[评估详情](#)，判断此训练作业是否满意。

训练作业详情

在ModelArts管理控制台，选择“训练作业”，进入训练作业列表页面。在训练作业列表中，您可以单击作业名称，查看该作业的详情。

每个版本的训练作业，包含的信息如**表5-21**所示。

图 5-9 训练作业详情

配置信息	日志	资源占用情况	评估结果
名称	ResNet_v1_50 job5f6f6746	算法	ResNet_v1_50
状态	运行成功	AI引擎	TensorFlow TF-1.8.0-python2.7
运行版本	V001	训练数据集	txfl_0530 V001
开始运行时间	2019/06/05 10:16:41	主要运行参数	-
运行时间	00:04:33	训练输出位置	/obs-lh/models/ResNet_v1_50/0605001/
资源池	2p100	描述	ased
计算节点个数	1	日志输出位置	-
NAS 挂载路径	-	NAS 地址	-

表 5-21 训练作业详情

参数	说明
V002	训练作业版本，由系统自动定义，命名规则为V001、V002。
状态	训练作业的状态。包含“部署中”、“运行中”、“运行成功”、“运行失败”、“已取消”。
运行时间	训练作业的运行时长。
配置信息	指当前训练作业版本的参数详情。
日志	指当前训练作业版本的运行日志。如果您在参数中设置了“作业日志路径”，您可以在“日志”页签单击“下载”将存储在OBS桶中的日志下载到本地。
资源占用情况	指当前训练作业版本的资源使用情况，资源包括CPU、GPU和内存。
评估结果	展示当前训练作业的评估结果，详细参数说明请参见 评估结果 。

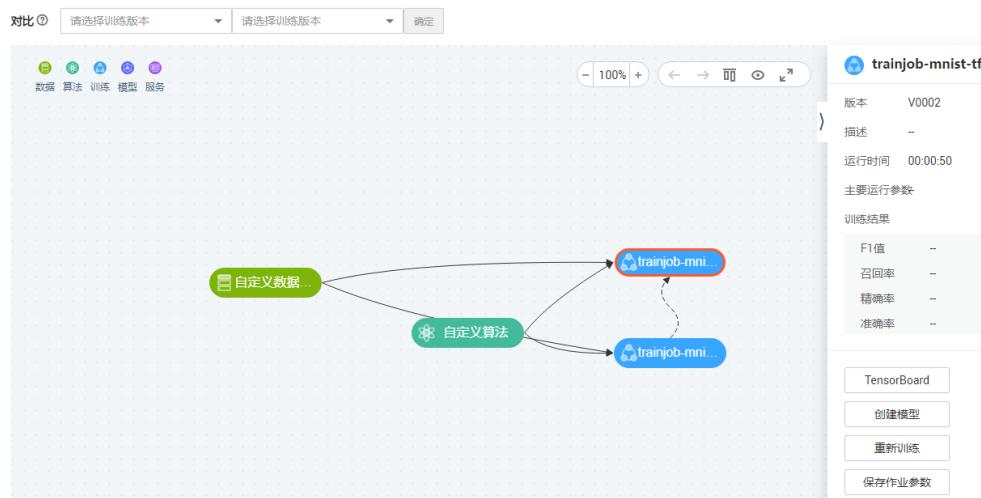
溯源图

在训练作业列表中，单击作业名称即可进入训练作业详情页面。默认进入“版本管理”页面，单击“溯源图”页签查看溯源图。

在“溯源图”页面中，您可以查看当前训练作业的数据、算法、训练、模型及服务之间的溯源图，并且您可以选择2个不同版本的训练作业进行对比。

在溯源图区域，选择任意一个元素，在界面右侧将展示此元素的详细信息，同时，还展示针对此元素您可以执行的下一步操作。例如当前选中的训练作业，右侧显示此训练作业的详细参数，并支持“创建TensorBoard”、“创建模型”、“重新训练”和“保存作业参数”的快捷操作。

图 5-10 查看溯源图



评估结果

在ModelArts管理控制台，选择“训练作业”，进入训练作业列表页面。训练作业运行成功后，您可以单击作业名称，进入“版本管理”页面，单击“评估结果”页签，查看训练作业的评估结果详情。包含标签列表、矩阵视图入口和自选矩阵视图入口，支持根据标签名称进行搜索查询。详细介绍请参见表5-22。

图 5-11 评估结果页

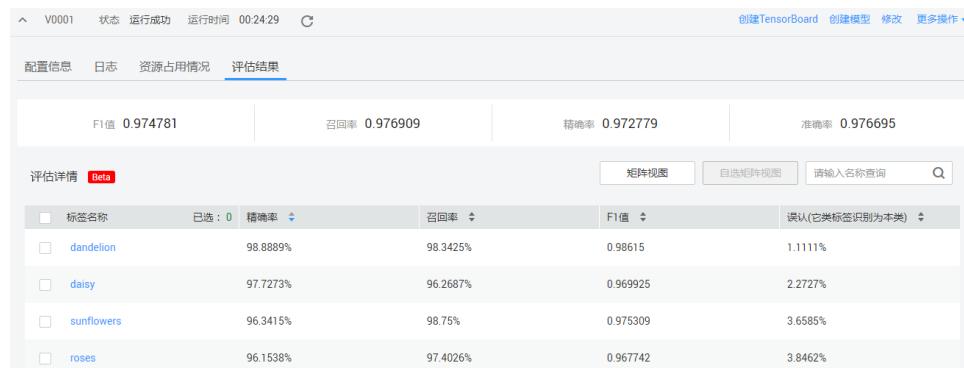


表 5-22 评估结果页内容介绍

功能模块	说明
标签列表	标签列表展示了“标签名称”、“精确率”、“召回率”、“F1值”和“误认”。
矩阵视图	单击“矩阵视图”，进入矩阵视图页面，可查看结果详情。默认展示矩阵视图。
自选矩阵视图	勾选目的标签，单击“自选矩阵视图”，查看目的标签结果详情。当无标签勾选时，“自选矩阵视图”为灰色状态。

标签详情

在“评估结果”页，单击标签名称查看标签预测结果。标签详情页列举了标签信息、预测结果、误认来源和矩阵视图界面入口，如图5-12所示。

图 5-12 标注详情页

The screenshot displays the 'Evaluation Results' page for the 'dandelion' label. Key performance metrics are shown at the top: Precision: 98.889%, Recall: 98.3425%, F1 Score: 0.98615, Error Rate: 1.1111%, and Total images: 180. Below these are sections for 'Predicted Results' and 'Misrecognition Sources'. The 'Predicted Results' section lists 'daisy' (3 images, 1.67%占比) and 'dandelion' (177 images, 98.33%占比). The 'Misrecognition Sources' section shows two images of flowers. Navigation controls like back, forward, and search are also present.

表 5-23 标签详情页内容介绍

功能模块	说明
标签信息	该标签对应的“精确率”、“召回率”、“F1值”、“误认率”和“本标签图片数”。
预测结果	预测的“标签名”、“图片数”和“占比”，支持图片数和占比的排序，按标签名搜索，预测结果可分页显示。
误认来源	误认为是该标签的图片列表，单击图片，展示该图片的标签信息及预测结果。
矩阵视图	单击标签名，进入矩阵视图界面。

矩阵视图

矩阵视图页列举了标签矩阵视图展示、矩阵网格缩放、筛选标签和预测特征可视化界面入口功能模块。

1. 在“评估结果”页，您可以通过如下方式查看矩阵视图。
 - 查看所有标签矩阵视图
 - 在“评估结果”页，单击页面右侧“矩阵视图”进入矩阵视图页面，查看所有标签的详细信息。
 - 在“评估结果”页，单击目标标签名称，进入标签详情界面，单击界面右侧“矩阵视图”，查看所选标签的详细信息。
 - 查看目标标签矩阵视图
 - 在“评估结果”页，勾选目标标签名称后，单击“自选矩阵视图”进入自选矩阵视图页面，查看所选标签的详细信息。
 - 在“评估结果”页，单击目标标签名称，在标签详情界面，单击“预测结果”下的标签名称，查看所选标签的详细信息。
2. 进入矩阵视图页面，您可以进行标签筛选、设置本页展示标签数量、查看数据和缩略图等操作。

图 5-13 矩阵视图页

预测结果		daisy	dandelion	roses	sunflowers																
标签		96.3% 129	1.5% 2	1.5% 2	0.7% 1																
daisy																					
dandelion		1.7% 3		98.3% 177																	
roses					97.4% 75	2.6% 2															
sunflowers						1.3% 1	98.8% 79														

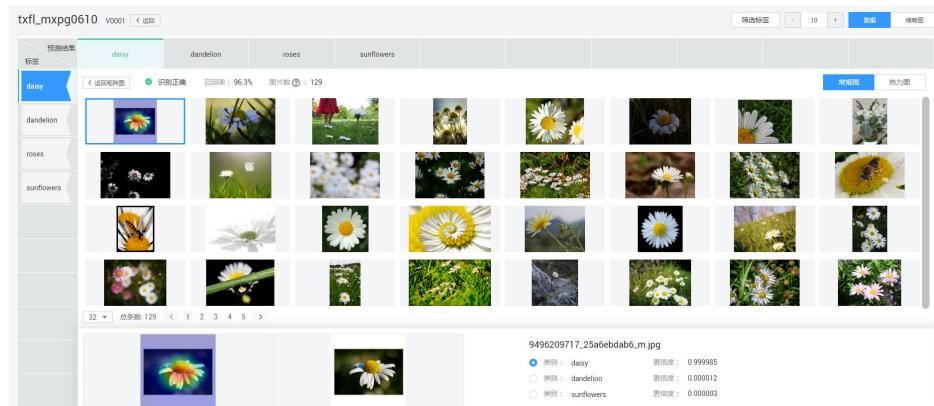
- 标签矩阵视图：标签矩阵视图展示分为“数据”和“缩略图”两种方式。纵坐标为标签，横坐标为预测结果，矩阵式呈现结果（召回率和图片数）。
- 筛选标签：单击右上角“筛选标签”，进入“筛选标签”页面。勾选对应标签和预测结果后，单击“确定”，返回矩阵视图页，通过筛选标签，可以改变矩阵图里“标签”和“预测结果”的关键词呈现量。
- 矩阵网格：矩阵网格的横纵坐标可展示规格为10*10、15*15、20*20的数据。

您可以通过右上角 10 区域进行操作。如果您筛选出来标签或预测结果的数大于可展示规格，那么您可以通过拉动右侧或下方的滚动条来查看其他标签或预测结果。

3. 在矩阵视图页面，可以在“数据”或“缩略图”页签单击不为空的网格，进入预测特征可视化界面。预测特征可视化界面分为标签预测基本信息、常规图和热力图列表、单张图片预测详情。

- 标签预测基本信息：标签预测，显示识别结果、召回率和图片数。
- 常规图和热力图列表：单击页面右侧“常规图”或“热力图”，通过选择标签和预测结果查看对应的常规图或热力图。
- 单张图片预测详情：单击选中单张图片，页面下方显示该图片常规图和热力图，展示预测类别和置信度。

图 5-14 预测特征可视化界面



4. 在预测特征可视化界面，如果图片数量过多，您可以单击页面下方的下拉列表 ，选择该页面展示的图片数量。

5.6 管理作业参数

创建训练作业时，您可以将训练作业的参数保存在ModelArts中，再次创建训练作业时，可一键使用已存储的作业参数，使得训练作业的创建高效便捷。

在创建训练作业、编辑训练作业、查看训练作业等操作过程中，保存的作业参数都将存储在“作业参数管理”页面中。

使用作业参数

- 方式1：在“作业参数管理”页面使用

登录ModelArts管理控制台，在左侧导航栏中选择“训练作业”，然后单击“作业参数管理”页签。在已有的作业参数列表中，单击“创建训练”，可快速将此作业参数用于创建一个新的训练作业。

- 方式2：在创建训练作业页面使用

在创建训练作业页面中，在“一键式参数配置参数”中，根据界面提示操作，选择需要使用的作业参数，快速创建一个可用的训练作业。

图 5-15 作业参数管理

The screenshot shows a table listing 13 job parameters. Each row contains the name, engine type, creation time, and a 'Description' column. A 'Operation' column on the right provides three actions: Create Training, Edit, and Delete. The table includes a search bar at the top and a footer with page navigation and total count information.

名称	引擎类型	创建时间	描述	操作
params-test-AA	TensorFlow	2018/12/03 16:41:43		创建训练 编辑 删除
Skllearn_1203	Scikit_Learn	2018/12/03 09:42:10	Skllearn_1203	创建训练 编辑 删除
aaaaaaaaaaaaaaa	MXNet	2018/11/30 22:14:21		创建训练 编辑 删除
Test_Custom	TensorFlow	2018/11/30 20:16:02	Test_Custom	创建训练 编辑 删除
Test_XGBoost_1130	XGBoost	2018/11/30 20:00:20	Test_XGBoost_1130	创建训练 编辑 删除
customimagedetect	MXNet	2018/11/30 19:41:10		创建训练 编辑 删除
ccf-113-01	TensorFlow	2018/11/30 19:37:39	ccf-113-01	创建训练 编辑 删除
Test_Test	TensorFlow	2018/11/23 11:09:27	Test_Test	创建训练 编辑 删除
FRR_v1_50_1122	TensorFlow	2018/11/22 23:05:50	FRR_v1_50_1122	创建训练 编辑 删除
XGboost_09	XGBoost	2018/11/22 22:48:11	XGboost_09	创建训练 编辑 删除

编辑作业参数

1. 登录ModelArts管理控制台，在左侧导航栏中选择“训练作业”，然后单击“作业参数管理”页签。
2. 在作业参数列表中，单击“操作”列的“编辑”。
3. 在打开的作业参数页面，参见**表5-17**，修改相关参数，然后单击“确定”保存此作业参数。

其中，作业参数的“名称”，不支持修改。

删除作业参数

1. 登录ModelArts管理控制台，在左侧导航栏中选择“训练作业”，然后单击“作业参数管理”页签。
2. 在作业参数列表中，单击“操作”列的“删除”。
3. 确认弹出对话框的信息，单击“确定”，完成删除操作。



说明

作业参数删除后不可恢复，请谨慎操作。

5.7 管理 TensorBoard

TensorBoard是一个可视化工具，能够有效地展示TensorFlow在运行过程中的计算图、各种指标随着时间的变化趋势以及训练中使用到的数据信息。TensorBoard当前只支持基于TensorFlow和MXNet引擎的训练作业。TensorBoard相关概念请参考[TensorBoard官网](#)。

对于采用AI引擎为TensorFlow和MXNet的训练作业，您可以使用模型训练时产生的Summary文件来创建TensorBoard作业。

前提条件

为了保证训练结果中输出Summary文件，在编写训练脚本时，您需要在代码中添加Summary相关代码。

- 使用TensorFlow引擎编写程序时

使用基于TensorFlow的MoXing时，需要将“mox.run”中设置参数“`save_summary_steps>0`”，并且超参“`summary_verbosity≥1`”。

如果您想显示其他指标，可以在“model_fn”的返回值类型“mox.ModelSpec”的“log_info”中添加张量（仅支持0阶张量，即标量），添加的张量会被写入到Summary文件中。如果您希望在Summary文件中写入更高阶的张量，只需要在“model_fn”中使用TensorFlow原生的“tf.summary”的方式添加即可。

- 使用MXNet引擎编写程序时

需要在代码里添加Summary相关代码，代码内容如下所示：

```
batch_end_callbacks.append(mx.contrib.tensorboard.LogMetricsCallback('s3路径'))
```

注意事项

- 运行中的TensorBoard会一直按需计费，当您不需要使用时，建议停止TensorBoard，避免产生不必要的费用。
- 默认使用CPU资源运行TensorBoard，且不支持修改为其他资源池。

创建 TensorBoard

1. 登录ModelArts管理控制台，在左侧导航栏中选择“训练作业”，然后单击“TensorBoard”页签。
2. 在TensorBoard列表中，单击左上方“创建”，进入创建TensorBoard界面。
3. 填写TensorBoard作业“名称”、“描述”，选择“日志路径”，其中“日志路径”选择创建训练作业时的“训练输出位置”。

图 5-16 创建 TensorBoard



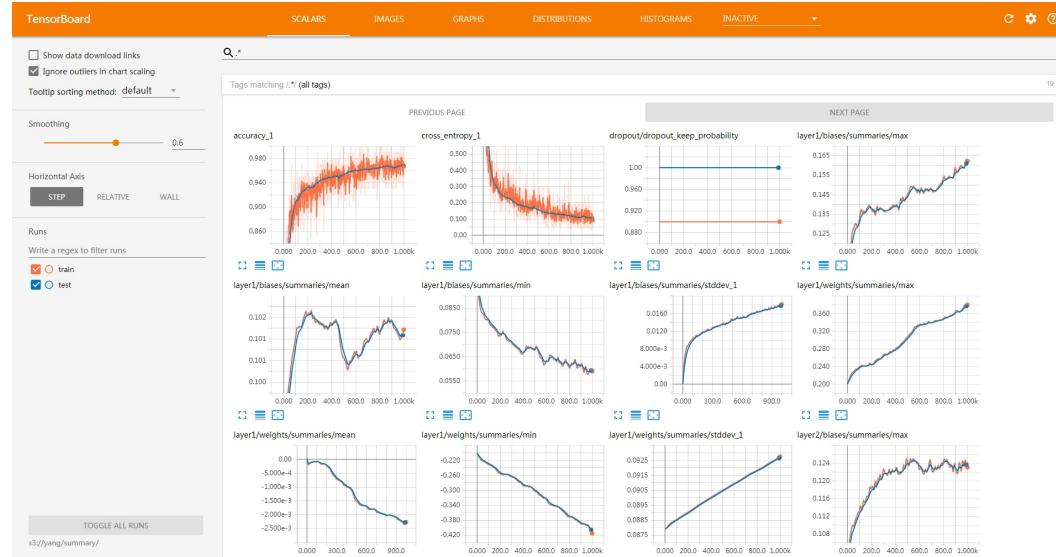
4. 参数填写完成后，单击“下一步”进行规格确认。
5. 规格确认无误后，单击“立即创建”，完成TensorBoard作业的创建。

在TensorBoard列表中，当状态变为“运行中”时，表示TensorBoard已创建完成。您可以单击TensorBoard名称进入查看详情。

打开TensorBoard

在TensorBoard列表中，单击TensorBoard作业名称，即可打开TensorBoard作业显示界面，如图5-17所示。只有“运行中”状态的TensorBoard支持打开。

图 5-17 TensorBoard 界面



运行或停止TensorBoard

- 停止TensorBoard:** 由于“运行中”的TensorBoard将一直按需计费，在不需要使用时，您可以停止TensorBoard停止计费。在TensorBoard列表中，单击“操作”列的“停止”，即可停止TensorBoard。
- 运行TensorBoard:** 对于“已取消”状态的TensorBoard，您可以重新运行并使用TensorBoard。在TensorBoard列表中，单击“操作”列的“运行”，即可运行TensorBoard。

删除TensorBoard

如果您的TensorBoard不再使用，您可以删除TensorBoard释放资源。在TensorBoard列表中，单击“操作”列的“删除”，即可删除TensorBoard。



说明

TensorBoard删除后不可恢复，需重新创建TensorBoard。请谨慎操作。

6 管理模型

6.1 模型管理简介

AI模型的开发和调优往往需要大量的迭代和调试，数据集、训练代码或参数的变化都可能会影响模型的质量，如不能统一管理开发流程元数据，可能会出现无法重现最优模型的现象。

ModelArts模型管理可导入所有训练版本生成的模型，可对所有迭代和调试的模型进行统一管理，通过数据集、训练和模型之间的版本演进溯源图，还可以实现模型的溯源管理。

模型管理支持的相关操作：

- [\(可选\) 购买模型调优](#)
- [导入模型](#)
- [管理模型版本](#)
- [模型二次调优](#)
- [将模型发布至市场](#)
- [模型模板](#)
- [转换模型](#)

6.2 (可选) 购买模型调优

ModelArts提供专业的模型调优服务，如果您对已有的模型不满意，且无法自行调整，您可以购买模型调优服务，由专业的工程师帮助您优化模型。

购买模型调优

1. 登录ModelArts管理控制台，在“总览”页面右侧的“费用”区域，单击“购买模型调优”。
2. 在“购买模型调优”页面，输入“付款金额”，然后在右下角勾选“我已阅读并同意《付款声明》”，然后单击“提交”。

提交完成后，华为云工程师将联系您了解具体需求。

图 6-1 购买模型调优

* 计费模式 一次性计费 按量计费

* 付款金额 万元
请填写付款金额，仅支持数字后一位小数点。

付款声明

用户须知：

1. 模型调优以收费方式提供。若您本人或他人（包括您的代理）通过个人帐户订阅收费服务，您应按照相关收费标准、付款方式支付相关服务费及其他费用。
2. 付款前请提前与客户经理确认您与华为已签订合同，并确认金额无误。付款及后续具体使用服务等情况，按照双方已签署的合同履行。

特此声明！

模型调优费用 **¥100,000.00**

我已阅读并同意《付款声明》

提交

6.3 导入模型

在模型训练完成后，其生成的模型存储在您指定的OBS目录下，您需要执行导入模型操作，将模型导入到ModelArts中进行管理。

为方便溯源和模型反复调优，在ModelArts中提供了模型版本管理的功能，您可以基于版本对模型进行管理。

使用前必读

- 如果使用ModelArts训练作业生成的模型，请确保训练作业已运行成功，且模型已存储至对应OBS目录下。
- 如果使用您本地编写好的模型，请确保您的模型符合ModelArts的规范要求，详细请参见[模型包规范介绍](#)。
- 对于模型首次上传至ModelArts时，请使用导入模型的功能，不能使用创建新版本功能。由于模型未导入，无法创建新版本。
- 自动学习项目中，在完成模型部署后，其生成的模型也将自动上传至模型管理列表中。但是自动学习生成的模型无法下载，只能用于部署上线。
- 导入和管理模型是免费的，不会产生费用。

导入模型

1. 登录ModelArts管理控制台，在左侧导航栏中选择“模型管理 > 模型列表”，进入模型列表页面。
2. 单击左上角的“导入”，进入“导入模型”页面。
3. 在“导入模型”页面，填写相关参数。
 - a. 填写模型基本信息，详细参数说明请参见[表6-1](#)。

表 6-1 模型基本信息参数说明

参数名称	说明
名称	模型名称。只支持1-48位可见字符（含中文），只能以大小写英文字母或中文字符开头，可包含字母、中文、数字、中划线、下划线。
版本	设置所创建模型的版本。第一次导入时，默认为0.0.1。
描述	模型的简要描述。

- b. 填写元模型来源及其相关参数。“元模型来源”有4种不同方式，请参见**表 6-2**选择。根据您选择的“元模型来源”不同，其相关的参数不同。

图 6-2 设置元模型来源及其相关参数



表 6-2 元模型来源参数说明

元数据来源	说明	相关的参数
从训练中选择	<p>从ModelArts已完成的训练作业中选择。</p> <ul style="list-style-type: none">当选择训练作业使用“预置模型”时，ModelArts默认提供推理代码和配置文件，无需提前上传。在此处，直接选择其对应的训练作业及版本即可。当选择的训练作业采用“常用框架”时，在导入模型之前，您需要按照模型包规范编写推理代码和配置文件，并将推理代码和配置文件放置元模型存储的“model”文件夹下。	<ul style="list-style-type: none">“推理代码”：显示模型推理代码URL，您可以直接复制此URL使用。“参数配置”：单击右侧的，查看当前模型的入参和出参。“运行时依赖”：罗列选中模型对环境的依赖。例如依赖“tensorflow”，安装方式为“pip”，其版本必须为1.8.0及以上版本

元数据来源	说明	相关的参数
从模板中选择	由于相同功能的模型配置信息重复率高，ModelArts将相同功能的配置整合成一个通用的模板，用户通过使用该模板，可以方便快捷的导入模型。模板的详细说明请参见 模板简介 。	<ul style="list-style-type: none">● “选择模板”：从已有的ModelArts模板列表中选择。例如，“TensorFlow图像分类模板”。● “模型目录”：指定模型存储的OBS路径。● “输入输出模式”：针对上方选择的模板，选择输入输出模式。所有输入输出模式的详细说明，请参见预置图像处理模式。
从OBS中选择	从OBS导入元模型。在“选择元模型”选择模型存储路径，此路径为训练作业中指定的“训练输出位置”。根据您选择的元模型存储路径，将自动关联出对应的“AI引擎”。 针对从OBS导入的元模型，ModelArts要求根据 模型包规范 ，编写推理代码和配置文件，并将推理代码和配置文件放置元模型存储的“model”文件夹下。如果您选择的目录下无对应的推理代码及配置文件，将无法导入模型。	<ul style="list-style-type: none">● “配置文件”：系统默认关联您存储在OBS中的配置文件。打开开关，您可以直接在当前界面查看、编辑或从OBS导入您的模型配置文件。● “参数配置”：单击右侧的，查看当前模型的入参和出参。● “运行时依赖”：罗列选中模型对环境的依赖。例如依赖“tensorflow”，安装方式为“pip”，其版本必须为1.8.0及以上版本
从容器镜像中选择	在“容器镜像所在的路径”右侧，单击  从容器镜像中导入模型的镜像，其中，模型均为Image类型，且不再需要用配置文件中的“swr_location”来指定您的镜像位置。 制作自定义镜像的操作指导及规范要求，请参见 自定义镜像简介 。 说明 您选择的模型镜像将共享给管理员，部署上线时，ModelArts将使用该镜像部署成推理服务，请确保您的镜像能正常启动并提供推理接口。	<ul style="list-style-type: none">● “配置文件”：支持“从OBS导入”或“在线编辑”的方式，配置文件需满足ModelArts编写规范，详情请参见模型包规范介绍。当选择“从OBS导入”时，您需要指定配置文件存储的OBS路径，且您可以打开“查看模型配置文件”右侧的开关，在线查看或编辑此配置文件。● “参数配置”：单击右侧的，查看当前模型的入参和出参。

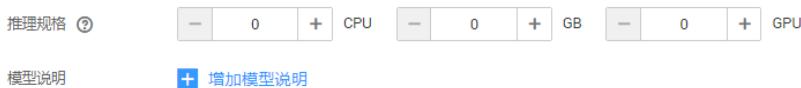
c. 设置最小推理规格和模型说明。

- “推理规格”：如果您的模型需要一定的规格资源才能完成推理，您可以配置推理规格，即您的模型部署上线进行正常推理所需要的最少规

格，后续的版本，部署上线阶段系统将会参考您填写的推理规格来分配资源。

- “模型说明”：为了帮助其他模型开发者更好的理解及使用您的模型，特别是需要会共享到模型市场的模型，建议您提供模型的说明文档。单击“增加模型说明”，设置“文档名称”及其“URL”。模型说明支持增加多条。

图 6-3 推理规格和模型说明



- 确认信息填写无误，单击“立即创建”，完成模型导入。

在模型列表中，您可以查看刚导入的模型及其对应的版本。当模型状态变更为“正常”时，表示模型导入成功。在此页面，您还可以创建新版本、快速部署模型、将模型发布至市场、导出模型、查看溯源图等操作。

6.4 管理模型版本

为方便溯源和模型反复调优，在ModelArts中提供了模型版本管理的功能，您可以基于版本对模型进行管理。

在模型版本管理页面，您还可以在操作列一键部署模型，部署模型的详细操作说明请参见[模型部署简介](#)。

前提条件

已在ModelArts中导入模型。且至少存在一个版本。

创建新版本

在“模型列表”页面，单击“创建新版本”进入“创建新版本”页面，参见[导入模型参数说明](#)填写相关参数，单击“立即创建”，完成新版本的创建操作。

查看溯源图

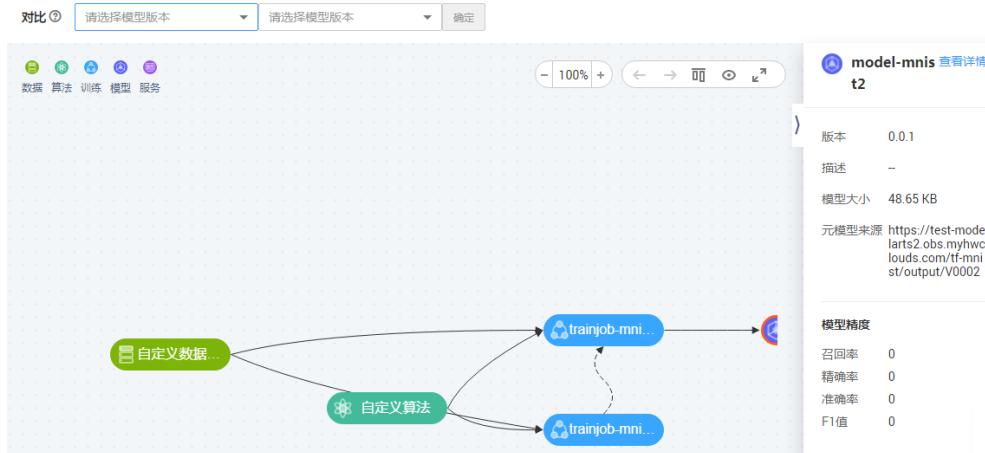
针对已导入的模型，您可以查看其溯源图，了解其数据、训练、模型等相关信息，方便您进行模型调优等操作。

在“模型管理 > 模型列表”页面，在左侧模型列表中选择一个模型，然后单击“溯源图”。在打开的溯源图页面，您可以查看到“数据”、“算法”、“训练”、“模型”及“服务”之间的溯源图，并且您可以选择2个不同版本的模型进行对比。

说明

如果您在导入模型时，“元数据来源”选择“从OBS中选择”或者“从容器镜像中选择”时，溯源图只能查看到“模型”的信息，前面的“数据”、“算法”、“训练”信息无法溯源。

图 6-4 查看溯源图



删除版本

在“模型管理 > 模型列表”页面，在版本列表中，单击“操作”列“更多>删除”，即可删除对应的版本。

说明

版本删除后不可恢复，请谨慎操作。

6.5 模型二次调优

模型二次调优即在原有模型的基础上，增加数据进行训练。目前只有“模型来源”为“预置算法”的模型支持二次调优。

前提条件

ModelArts模型管理中已存在“模型来源”为“预置算法”的模型。

操作步骤

1. 登录ModelArts管理控制台，在左侧导航栏中选择“模型管理 > 模型列表”，进入模型列表页面。
2. 选择“模型来源”为“预置算法”的模型名称，单击操作列的“二次调优”，进入“创建训练作业”页面。
3. 在“创建训练作业”页面，填写相关参数。具体操作及参数说明请参见[创建训练作业](#)。
4. 训练作业运行成功后，会在“模型管理”页面，目标模型栏目下，生成调优后新的模型版本，该模型描述为“finetuned from version: v x.x.x”。至此即完成一次模型调优。

6.6 将模型发布至市场

ModelArts提供了“AI市场”功能，方便将个人的模型、API、数据集等共享给所有ModelArts用户，您也可以从“AI市场”获取他人共享的内容，快速完成构建。在您完成模型的训练和导入之后，您可以将自己的模型分享至“AI市场”，进行知识共享。

前提条件

已在ModelArts中导入模型。且至少存在一个版本。

操作步骤

1. 登录ModelArts管理控制台，在左侧导航栏中选择“模型管理 > 模型列表”，进入模型列表页面。
2. 在版本列表中，单击“操作”列的“更多>市场发布”。
3. 在弹出的对话框中填写参数，参数填写如表6-3所示。

图 6-5 发布模型

发布模型

* 发布者

* 名称

描述

模型画像

行业	<input type="text"/>
数据	<input type="text"/>
场景	<input type="text"/>
主题	<input type="text"/>
模型	<input type="text"/>
框架引擎	<input type="text"/>

* 封面图 使用上一次的封面图 OBS 本地上传

* 发布到 AI市场 个人 [如何查看账号id](#)

已发布给0个账号

账号	操作
	<input type="button" value="确定"/> <input type="button" value="取消"/>

表 6-3 参数说明

参数名称	说明
发布者	模型市场中显示的发布者名称，发布后将不能修改。
名称	模型市场中显示的模型名称。

参数名称	说明
描述	对模型的简要描述，建议从使用场景、使用方法、训练数据集三个方面描述您的模型。
模型画像	设置模型画像后，可在市场中显示模型的标签，便于分类显示和快速查找到模型。 建议从每个属性的菜单中选择至多三个最合适的标签用于描述您的模型，如没有合适的标签，也可留空不选。可以从行业、数据、场景、主题、模型、框架引擎几个类型中选择标签。
封面图	封面图可帮助其他用户直观感受您的模型的用途，您可以从“OBS”中选取，有可以从“本地上传”，如果您已共享过模型也可以使用上一次的封面图。 GIF动态图会是非常好的选择，支持jpg、png、gif、bmp格式的图像，最佳宽高比是“5:3”。
发布到	可共享到“AI市场”，也可共享给“个人”。 <ul style="list-style-type: none">● “AI市场”：共享给所有用户。共享到AI市场需要审核，可在共享对话框中查看审批状态。● “个人”：共享给指定用户。 说明<ul style="list-style-type: none">● 账号ID可从“我的凭证”中获取。● 输入多个用户ID时，用户ID之间请用“，”隔开，且不允许出现特殊字符及空格。

4. 单击“确定”，完成模型的发布操作。

您可以在ModelArts管理控制台“AI市场”中“我发布的”页面查看您的发布，查看您的发布操作请参见[我发布的](#)。

6.7 模型模板

6.7.1 模板简介

背景信息

相同功能的模型配置信息重复率高，将相同功能的配置整合成一个通用的模板，用户通过使用该模板，可以方便快捷的导入模型，而不用编写配置文件。简单来说，模板将AI引擎以及模型配置模板化，每种模板对应于1种具体的AI引擎及1种推理模式，借助模板，用户可以快速导入模型到ModelArts。

模板分两大类型：通用类型，非通用类型。

- 非通用类型模板，针对特定的场景所定制的，固定输入输出模式，不可覆盖，如“TensorFlow图像分类模板”，固定使用预置图像处理模式。
- 通用模板，搭载特定的AI引擎以及运行环境，内置的输入输出模式为未定义模式，即不定义具体的输入输出格式，用户需根据模型功能或业务场景重新选择新的输入输出模式来覆盖内置的未定义模式，如图像分类模型应选择预置图像处理模式，而目标检测模型则应选择预置物体检测模式。

说明

使用未定义模式的模型将无法部署批量服务。

如何使用模板

以“TensorFlow图像分类模板”为例，您需要事先将TensorFlow模型包上传至OBS，模型包结构请参见[模板模型包规范](#)。模型文件应存放在model目录下，通过该模板创建模型时，您需要选择到model这一目录。

1. 在“导入模型”页面，在“元模型来源”参数中选择“从模板中选择”。
2. 在“选择模板”区域，选择“TensorFlow图像分类模板”。

ModelArts还提供“类型”、“引擎”、“环境”三个筛选条件，帮助您更快找到想要的模板。如果这个三个筛选条件不能满足您的要求，可以使用关键词搜索，找到目标模板。

图 6-6 选择模板



3. 在“模型目录”中，选择模型文件存储的model目录。详细规范请参见[模板模型包规范](#)。

说明

当训练作业执行多次时，将基于V001、V002等规则生成不同的版本目录，且生成的模型将存储在不同版本目录下的model文件夹。此处选择模型文件时，需指定对应版本目录下的model文件夹。

图 6-7 设置模型目录



4. 如果您选择的模板允许覆盖其中的默认输入输出模式，您可以根据模型功能或业务场景在“输入输出模式”中，选择相应的输入输出模式。“输入输出模式”是对“config.json”中API的抽象，描述模型对外提供推理的接口。一个“输入输出模式”描述一个或多个API接口，每个模板对应于一个“输入输出模式”。

例如，“TensorFlow图像分类模板”，其支持的“输入输出模式”为“预置图像处理模式”，但该模板不允许修改其中的输入输出模式，所以您在页面上只能看到模板默认的输入输出模式，而不能选择其他模式。

当前支持的“输入输出模式”及其说明请参见[输入输出模式说明](#)。

当前支持的模板

表 6-4 ModelArts 提供的模板

模板名称	描述	对应的输入输出模式
“TensorFlow图像分类模板”	搭载TensorFlow引擎，运行环境为“python2.7”。该模板使用平台预置的图像处理模式，使用该模板导入的模型推理时，您需要采用“multipart/form-data”格式发送一张“key”为“images”的图像文件（Content-type:multipart/form-data、key:images、type:file）。	预置图像处理模式(不可覆盖)
“TensorFlow-py27通用模板”	搭载TensorFlow1.8 AI引擎，运行环境为“python2.7”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)
“TensorFlow-py36通用模板”	搭载TensorFlow1.8 AI引擎，运行环境为“python3.6”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)
“MXNet-py27通用模板”	搭载MXNet1.2.1 AI引擎，运行环境为“python2.7”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)
“MXNet-py36通用模板”	搭载MXNet1.2.1 AI引擎，运行环境为“python3.6”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)
“PyTorch-py27通用模板”	搭载PyTorch1.0 AI引擎，运行环境为“python2.7”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)

模板名称	描述	对应的输入输出模式
“PyTorch-py36通用模板”	搭载PyTorch1.0 AI引擎，运行环境为“python3.6”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)
“Caffe-CPU-py27通用模板”	搭载Caffe1.0 CPU版 AI引擎，运行环境为“python2.7”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录那一层。	未定义模式(可覆盖)
“Caffe-GPU-py27通用模板”	搭载Caffe1.0 GPU版 AI引擎，运行环境为“python2.7”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)
“Caffe-CPU-py36通用模板”	搭载Caffe1.0 CPU版 AI引擎，运行环境为“python3.6”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)
“Caffe-GPU-py36通用模板”	搭载Caffe1.0 GPU版 AI引擎，运行环境为“python3.6”，内置输入输出模式为未定义模式，请根据模型功能或业务场景重新选择合适的输入输出模式。使用该模板导入模型时请选择到包含模型文件的model目录。	未定义模式(可覆盖)

6.7.2 模板模型包规范

在使用模型模板时，您选择的模型目录需满足ModelArts定义的规范，详细要求请参见[模型包规范](#)。针对不同的引擎框架，其对应的目录结构存在一些区别，详情可参考[模型包示例](#)。

模型包规范

- 模型包必须存储在OBS中，且必须以“model”命名。“model”文件夹下面放置模型文件、模型推理代码。
- 模型推理代码文件不是必选文件，如果有，其文件名必须为“customize_service.py”，“model”文件夹下有且只能有1个推理代码文件，模型推理代码编写请参见[模型推理代码编写说明](#)。
- 使用模板导入的模型包结构如下所示：

```
model/
  |
  └── 模型文件          //必选，不同的框架，其模型文件格式不同，详细可参考模型包示例。
  └── 自定义Python包    //可选，用户自有的Python包，在模型推理代码中可以直接引用。
```

|—— customize_service.py //可选，模型推理代码，文件名称必须为“customize_service.py”，否则不视为推理代码。

模型包示例

● TensorFlow模型包结构

发布该模型时只需要指定到“model”目录。

OBS桶/目录名

```
|—— model    必选，文件夹名称必须为“model”，用于放置模型相关文件。  
|   |—— <<自定义python包>>    可选，用户自有的Python包，在模型推理代码中可以直接引用。  
|   |—— saved_model.pb        必选，protocol buffer格式文件，包含该模型的图描述。  
|   |—— variables            对“*.pb”模型主文件而言必选。文件夹名称必须为“variables”，包含模型的权重偏差等信息。  
|   |   |—— variables.index      必选  
|   |   |—— variables.data-00000-of-00001    必选  
|   |—— customize_service.py    可选，模型推理代码，文件名称必须为“customize_service.py”，有且只有1个推理代码文件。“customize_service.py”依赖的“py”文件可以直接放“model”目录下。
```

● MXNet模型包结构

发布该模型时只需要指定到“model”目录。

OBS桶/目录名

```
|—— model    必选，文件夹名称必须为“model”，用于放置模型相关文件。  
|   |—— <<自定义python包>>    可选，用户自有的Python包，在模型推理代码中可以直接引用。  
|   |—— resnet-50-symbol.json    必选，模型定义文件，包含模型的神经网络描述。  
|   |—— resnet-50-0000.params    必选，模型变量参数文件，包含参数和权重信息。  
|   |—— customize_service.py    可选，模型推理代码，文件名称必须为“customize_service.py”，有且只有1个推理代码文件。“customize_service.py”依赖的“py”文件可以直接放“model”目录下。
```

● pyspark模型包结构

发布该模型时只需要指定到“model”目录。

OBS桶/目录名

```
|—— model    必选，文件夹名称必须为“model”，用于放置模型相关文件。  
|   |—— <<自定义Python包>>    可选，用户自有的Python包，在模型推理代码中可以直接引用。  
|   |—— spark_model        必选，模型文件夹，包含pyspark保存的模型内容。  
|   |—— customize_service.py    可选，模型推理代码，文件名称必须为“customize_service.py”，有且只有1个推理代码文件。“customize_service.py”依赖的“py”文件可以直接放“model”目录下。
```

● PyTorch模型包结构

发布该模型时只需要指定到“model”目录。

OBS桶/目录名

```
|—— model    必选，文件夹名称必须为“model”，用于放置模型相关文件。  
|   |—— <<自定义Python包>>    可选，用户自有的Python包，在模型推理代码中可以直接引用。  
|   |—— resnet50.pth        必选，pytorch模型保存文件，存有权重变量等信息。  
|   |—— customize_service.py    可选，模型推理代码，文件名称必须为“customize_service.py”，有且只有1个推理代码文件。“customize_service.py”依赖的“py”文件可以直接放“model”目录下。
```

● Caffe模型包结构

发布该模型时只需要指定到“model”目录。

OBS桶/目录名

```
|—— model    必选，文件夹名称必须为“model”，用于放置模型相关文件。  
|   |—— <<自定义python包>>    可选，用户自有的Python包，在模型推理代码中可以直接引用。  
|   |—— deploy.prototxt        必选，caffe模型保存文件，存有模型网络结构等信息。  
|   |—— resnet.caffemodel      必选，caffe模型保存文件，存有权重变量等信息。  
|   |—— customize_service.py    可选，模型推理代码，文件名称必须为“customize_service.py”，有且只有1个推理代码文件。“customize_service.py”依赖的“py”文件可以直接放“model”目录下。
```

6.7.3 输入输出模式说明

6.7.3.1 预置物体检测模式

输入

系统预置物体检测输入输出模式，预测请求路径“/”，请求协议为“HTTP”，请求方法为“POST”，调用方需采用“multipart/form-data”内容类型，以“key”为“images”，“type”为“file”的格式输入待处理图片。

输出

推理结果以“JSON”体的形式返回，具体字段请参见[表6-5](#)。

表 6-5 参数说明

字段名	类型	描述
detection_classes	字符串数组	输出物体的检测类别列表，如["yunbao","cat"]
detection_boxes	数组，元素为浮点数数组	输出物体的检测框坐标列表，坐标表示为[Ymin,Xmin,Ymax,Xmax]
detection_scores	浮点数数组	输出每种检测列表的置信度，用来衡量识别的准确度。

推理结果的“JSON Schema”表示如下：

```
{  
    "type": "object",  
    "properties": {  
        "detection_classes": {  
            "item": {  
                "type": "string"  
            },  
            "type": "array"  
        },  
        "detection_boxes": {  
            "items": {  
                "minItems": 4,  
                "items": {  
                    "type": "number"  
                },  
                "type": "array",  
                "maxItems": 4  
            },  
            "type": "array"  
        },  
        "detection_scores": {  
            "item": {  
                "type": "string"  
            },  
            "type": "array"  
        }  
    }  
}
```

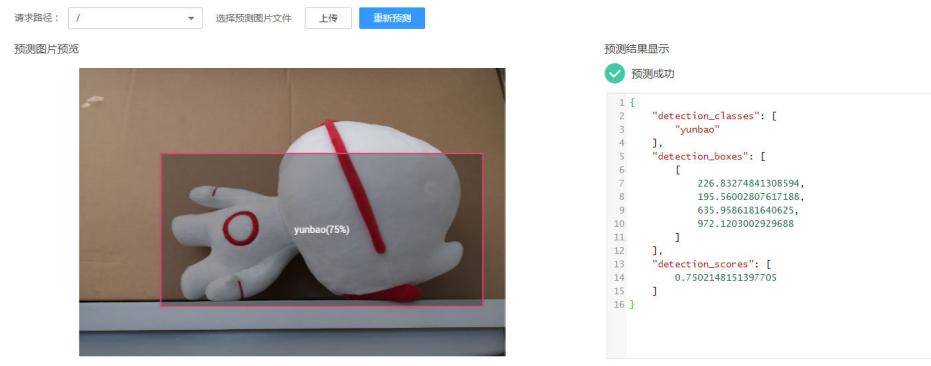
请求样例

该模式下的推理方式均为输入一张待处理图片，推理结果以“JSON”格式返回。示例如下：

- 页面预测

在服务详情的“预测”页签，上传需要检测的图片，单击“预测”即可获取检测结果。

图 6-8 页面预测



- Postman调REST接口预测

部署上线成功后，您可以从服务详情页的调用指南中获取预测接口地址，预测步骤如下：

- 选择“Headers”设置请求头部，“Content-Type”的值设为“multipart/form-data”，“X-Auth-Token”的值设为用户实际获取的token值。

图 6-9 设置请求头部

KEY	VALUE	DESCRIPTION
Content-Type	multipart/form-data	
X-Auth-Token	{{token}}	

- 选择“Body”设置请求体，“key”选择为“images”，选择为“File”类型，接着通过选择文件按钮选择需要处理的图片，最后单击“send”，发送您的预测请求。

图 6-10 设置请求体

KEY	VALUE	DESCRIPTION
images	File	IMG_2018091...

This file resides outside the working directory. Collaborators might not have the same file path.

```
1 {
2     "detection_classes": [
3         "yunbao"
4     ],
5     "detection_boxes": [
6         [
7             376.0593566894531,
8             68.6734390258789,
9             1191.673005703125,
10            1012.961669921875
11        ],
12    ],
13    "detection_scores": [
14        0.7887529134750366
15    ]
16 }
```

6.7.3.2 预置图像处理模式

输入

系统预置图像处理输入输出模式，预测请求路径“/”，请求协议为“HTTPS”，请求方法为“POST”，调用方需采用“multipart/form-data”内容类型，以“key”为“images”，“type”为“file”的格式输入待处理图片。

输出

推理结果以“JSON”体的形式返回，“JSON”的具体字段由模型决定。

请求样例

该模式下的推理方式均为输入一张待处理图片，响应的“JSON”根据模型改变而改变。示例如下：

- 页面预测

图 6-11 在界面中直接预测



- Postman 调 REST 接口预测

部署上线成功后，您可以从服务详情页的调用指南中获取预测接口地址。选择“Body”设置请求体，“key”选择为“images”，选择为“File”类型，接着通过选择文件按钮选择需要处理的图片，最后单击“send”，发送您的预测请求。

图 6-12 调用 REST 接口



6.7.3.3 预置预测分析模式

输入

系统预置预测分析输入输出模式，预测请求路径“/”，请求协议为“HTTP”，请求方法为“POST”，调用方需采用“application/json”内容类型，发送预测请求，请求体以“JSON”格式表示，“JSON”字段说明请参见表6-6。

表 6-6 JSON 字段说明

字段名	类型	描述
data	Data结构	包含预测数据。“Data结构”说明请参见 表 6-7 。

表 6-7 Data 结构说明

字段名	类型	描述
req_data	ReqData结构数组	预测数据列表。

“ReqData”，是“Object”类型，表示预测数据，数据的具体组成结构由业务场景决定。使用该模式的模型，其自定义的推理代码中的预处理逻辑应能正确处理模式所定义的输入数据格式。

预测请求的“JSON Schema”表示如下：

```
{  
    "type": "object",  
    "properties": {  
        "data": {  
            "type": "object",  
            "properties": {  
                "req_data": {  
                    "items": [{  
                        "type": "object",  
                        "properties": {}  
                    }],  
                    "type": "array"  
                }  
            }  
        }  
    }  
}
```

输出

预测结果以“JSON”格式返回，“JSON”字段说明请参见[表6-8](#)。

表 6-8 JSON 字段说明

字段名	类型	描述
data	Data结构	包含预测数据。“Data结构”说明请参见 表 6-9 。

表 6-9 Data 结构说明

字段名	类型	描述
resp_data	RespData结构数组	预测结果列表。

与“ReqData”一样，“RespData”也是“Object”类型，表示预测结果，其具体组成结构由业务场景决定。我们建议使用该模式的模型，其自定义的推理代码中的后处理逻辑应输出符合模式所定义的数据。

预测结果的“JSON Schema”表示如下：

```
{  
    "type": "object",  
    "properties": {  
        "data": {  
            "type": "object",  
            "properties": {  
                "resp_data": {  
                    "type": "array",  
                    "items": [{  
                        "type": "object",  
                        "properties": {}  
                    }]  
                }  
            }  
        }  
    }  
}
```

请求样例

该模式下的推理方式均为输入“JSON”格式的待预测数据，预测结果以“JSON”格式返回。示例如下：

- 页面预测

在服务详情的“预测”页签，输入预测代码，单击“预测”即可获取检测结果。

图 6-13 页面预测

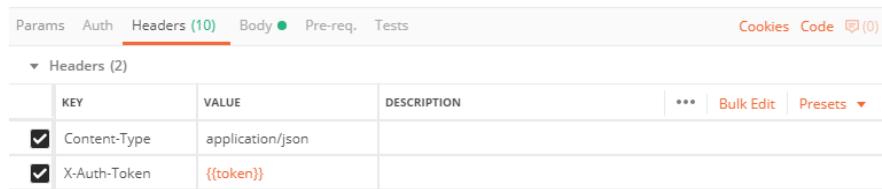


- Postman调REST接口预测

部署上线成功后，您可以从服务详情页的调用指南中获取预测接口地址，预测步骤如下：

- 选择“Headers”设置请求头部，“Content-Type”的值设为“application/json”，“X-Auth-Token”的值设为用户实际获取的token值。

图 6-14 设置请求头部



- 选择“Body”设置请求体，编辑需要预测的数据，最后单击“send”，发送您的预测请求。

6.7.3.4 未定义模式

描述

未定义的模式，即不定义具体的输入输出格式，请求的输入输出完全由模型决定。当现有的输入输出模式不适合模型的场景时，才考虑选择该模式。使用未定义模式导入的模型无法部署批量服务，同时服务的预测界面可能无法正常工作。我们会不断提供新模式，力求覆盖更多的场景，让每个模型都有合适的模式可选择。

输入

不限。

输出

不限。

请求样例

未定义模式没有特定的请求样例，请求的输入输出完全由模型决定。

6.8 转换模型

6.8.1 转换操作

针对您在ModelArts或者本地构建的模型，为获得更高的算力，希望将模型应用于Ascend芯片、ARM或GPU上，此时，您需要将已有模型转换成相应的格式后，再应用至不同的芯片类型。

ModelArts提供了模型转换功能，即将已有的模型转换成所需格式，以便应用于算力和性能更高的芯片上。

模型转换主要应用场景如下所示：

- 使用Caffe（.caffemodel格式）或者Tensorflow框架（frozen_graph格式）训练模型，使用转换功能可将模型转换成om格式，转换后的模型可华为昇腾（Ascend）芯片上部署运行
- 使用Tensorflow框架训练模型（frozen_graph或“saved_model”格式），使用转换功能可以将模型转换量化成tflite格式模型，转换后的模型可以在ARM上部署运行
- 使用Tensorflow框架训练模型（frozen_graph或“saved_model”格式），使用转换功能可以将模型转换量化成tensorRT格式模型，转换后的模型可以在nvidia p4 GPU上部署运行

模型转换说明

- 模型转换当前只支持三种芯片类型，分别为：Ascend、ARM、GPU。
- 模型转换当前仅支持Caffe和TensorFlow框架训练输出的模型。

- ModelArts提供了转换模板供用户选择，只能选择对应模板进行转换，支持的模板描述，请参见[转换模板](#)。
- 现阶段由于tflite和tensorRT支持的算子和量化算子有限，可能存在部分模型转换失败的情况，如果出现转换失败，可以通过提[工单](#)获得专业的技术支持。

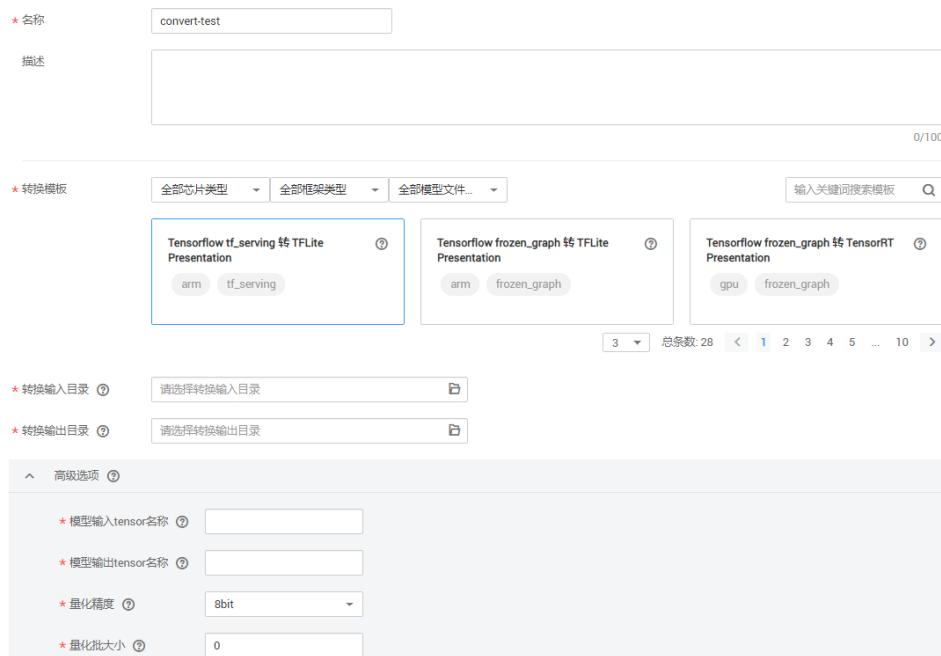
创建模型转换任务

1. 登录ModelArts管理控制台，在左侧导航栏中选择“模型管理 > 模型转换列表”，进入模型转换列表页面。
2. 单击左上角的“创建模型转换任务”，进入任务创建页面。
3. 在“创建模型转换任务”页面，参考[表6-10](#)，填写相关信息。

表 6-10 参数说明

参数	说明
名称	转换任务的名称。
描述	转换任务的简要描述。
转换模板	<p>ModelArts提供了一系列的模板，定义转换功能以及转换过程中所需的参数。</p> <p>当前支持的转换模板详细描述请参见转换模板。您可以从界面的模板卡片列表中选择所需模板。您也可以通过右侧搜索框输入关键词搜索模板，或者基于“芯片类型”、“框架类型”、“模型文件格式”三个维度筛选所需模板。</p> <ul style="list-style-type: none">● “芯片类型”：ModelArts转换模板支持三种芯片类型，分别为Ascend、ARM、GPU。● “框架类型”：转换模板基于不同的框架，生成不同格式的模型。目前支持Tensorflow和Caffe两种框架类型。● “模型文件格式”：下拉列表罗列当前支持的模型文件格式，通过模型文件格式选择。目前支持“caffemodel”、“frozen_gragh”、“tf_serving”文件格式。
转换输入目录	用于转换的模型所在目录，此目录必须为OBS目录，且模型文件的目录需符合ModelArts规范，详情请参见 模型输入目录规范 。
转换输出目录	模型转换完成后，根据此参数设置的目录存储模型。输出目录需符合ModelArts规范要求，详情请参见 模型输出目录说明 。
高级选项	<p>针对不同的转换模板，ModelArts提供了更多的选项设置，例如量化精度，方便您可以对模型转换任务进行更高阶的设置。</p> <p>不同的转换模板，其对应的高级选项支持的参数不同，每个模板支持的详细参数，请参见转换模板。</p>

图 6-15 创建模型转换任务



4. 任务信息填写完成后，单击右下角“立即创建”。

创建完成后，系统自动跳转至“模型转换列表”中。刚创建的转换任务将呈现在界面中，其“任务状态”为“初始化”。任务执行过程预计需要几分钟到十几分钟不等，请耐心等待，当“任务状态”变为“成功”时，表示任务运行完成并且模型转换成功。

如果“任务状态”变为“失败”，建议单击任务名称进入详情页面，查看日志信息，根据日志信息调整任务的相关参数并创建新的转换任务。

删除转换任务

针对运行结束的任务，如果不需要再使用，您可以删除转换任务。其中，“运行中”或“初始化”状态中的任务不支持删除操作。

说明

任务删除后，将无法恢复，请谨慎操作。

● 删除单个

在“模型转换列表”中，针对需要删除的单个任务，您可以在此任务所在行，单击操作列的“删除”，完成删除操作。

● 批量删除：

在“模型转换列表”中，勾选多个待删除的任务，然后单击左上角“删除”，完成批量任务的删除操作。

6.8.2 模型输入目录规范

模型转换后，应用于不同的芯片，针对不同的芯片，其模型输入目录的要求不同。ModelArts当前对模型输入目录的要求分为**Ascend芯片**和**ARM或GPU**两种。

Ascend 芯片

用于Ascend芯片的模型，其转换要求如下所示：

- 针对基于Caffe框架的模型，执行模型转换时，其输入目录需符合如下规范。

---xxxx.caffemodel	模型参数文件，输入目录下有且只能有一个。
---xxxx.prototxt	模型网络文件，输入目录下有且只能有一个。
---insert_op_conf.cfg	插入算子配置文件，输入目录下有且还有一个。
---plugin	自定义算子目录，可以没有。如果有，输入目录下有且只能有一个plugin文件夹。

- 针对基于TensorFlow框架的模型，执行模型转换时，其输入目录需符合如下规范。

---xxxx.pb	模型网络文件，输入目录下有且只能有一个。当前只支持以frozen_graph保存的模型。
---insert_op_conf.cfg	插入算子配置文件，输入目录下有且还有一个。
---plugin	自定义算子目录，可以没有。如果有，输入目录下有且只能有一个plugin文件夹。

ARM 或 GPU

用于ARM或GPU的模型，当前只支持TensorFlow框架的模型，包含两种格式“frozen_graph”和“saved_model”。

“frozen_graph”格式如下所示：

---model	模型存放目录，必须以model命名，有且只能有一个，目录下只能放一个模型相关文件。
---xxx.pb	模型文件。必须是tensorflow的frozen_graph格式的文件。
---calibration_data	校准数据集存放目录，必须以calibration_data命名，8bit转换需要，32bit转换不需要。输入目录下有且只能有一个。
---xx.npy	校准数据集。可以是多个npy格式文件，需要确保npy是在预处理后直接输入模型的数据，其输入的tensor需要与模型输入保持一致。

“saved_model”格式如下所示：

---model	模型存放目录，必须以model命名，有且只能有一个，目录下只能放一个模型相关文件。
---saved_model.pb	模型文件。必须是tensorflow的saved_model格式的文件。
---variables	变量存储文件夹。
---variables.data-*****-of-*****	saved_model格式文件需要的数据。
---variables.index	saved_model格式文件需要的索引。
---calibration_data	校准数据集存放目录，必须以calibration_data命名，8bit转换需要，32bit转换不需要。输入目录下有且只能有一个。
---xx.npy	校准数据集。可以是多个npy格式文件，需要确保npy是在预处理后直接输入模型的数据，其输入的tensor需要与模型输入保持一致。

6.8.3 模型输出目录说明

转换模型任务执行完成后，ModelArts将转换后的模型输出至指定的OBS路径。针对不同的转换任务，基于不同的芯片，其对应的目录有所区别，分为Ascend芯片和ARM或GPU两种。

Ascend 芯片

用于Ascend芯片的模型，其转换后输出目录说明如下所示：

- 针对基于Caffe框架的模型，执行模型转换时，其输出目录说明如下所示。

---xxxx.om	转换输出的模型，可用于Ascend芯片，模型文件后缀统一为“.om”。
---job_log.txt	转换过程的日志文件。

- 针对基于TensorFlow框架的模型，执行模型转换时，其输出目录说明如下所示。

---xxxx.om	转换输出的模型，可用于Ascend芯片，模型文件后缀统一为“.om”。
---job_log.txt	转换过程的日志文件。

ARM 或 GPU

用于ARM或GPU的模型，其转换后输出目录说明如下所示：

GPU格式如下所示：

```
|---model
|   |---xxx.pb          GPU转换后模型后缀为“.pb”。
|---job_log.txt         转换过程的日志文件。
```

ARM格式如下所示：

```
|---model
|   |---xxx.tflite      ARM转换后模型后缀为“.tflite”。
|   |---config.json     8bit转换后，用户需要使用tflite时需要的参数。
|---job_log.txt        转换过程的日志文件。
```

6.8.4 转换模板

表 6-11 ModelArts 提供转换模板

模板名称	模板描述	模板高级选项
Caffe转Ascend	转换Caffe框架训练出来的模型，转换后模型可在Ascend芯片上运行。	无
Tensorflow frozen_graph转TFLite	转换Tensorflow框架训练并以“frozen_graph”格式保存的模型，转换后模型可在ARM上运行。	<ul style="list-style-type: none">● 模型输入tensor名称：以字符串形式输入模型输入张量名称，以“input1:input2”形式表示。● 模型输出tensor名称：以字符串形式输入模型输出张量名称，以“output1:output2”形式表示。● 量化精度：可选择8bit或32bit。32bit表示直接转换模型，8bit表示模型进行量化。● 量化批大小：以数值形式输入量化批大小。必须为正整数。
Tensorflow saved_model转TFLite	转换Tensorflow框架训练并以“saved_model”格式保存的模型，转换后模型可在ARM上运行。	<ul style="list-style-type: none">● 模型签名：以字符串形式输入模型输入tensor签名，默认会选择第一个签名。● 传入模型标签：以字符串形式输入模型输出标签，默认会选择第一个标签。● 量化精度：可选择8bit或32bit。32bit表示直接转换模型，8bit表示模型进行量化。● 量化批大小：以数值形式输入量化批大小。必须为正整数。

模板名称	模板描述	模板高级选项
Tensorflow frozen_graph转 TensorRT	转换Tensorflow框架训练并以 “frozen_graph”格式 保存的模型，转换后 模型可在GPU上运行。	<ul style="list-style-type: none">● 模型输入tensor名称：以字符串形式输入模型输入张量名称，以“input1:input2”形式表示。● 模型输出tensor名称：以字符串形式输入模型输出张量名称，以“output1:output2”形式表示。● 量化精度：可选择8bit或32bit。32bit表示直接转换模型，8bit表示模型进行量化。● 量化批大小：以数值形式输入量化批大小。必须为正整数。
Tensorflow saved_model转 TensorRT	转换Tensorflow框架训练并以 “saved_model”格式 保存的模型，转换后 模型可在GPU上运行。	<ul style="list-style-type: none">● 模型签名：以字符串形式输入模型输入tensor签名，默认会选择第一个签名。● 传入模型标签：以字符串形式输入模型输出标签，默认会选择第一个标签。● 量化精度：可选择8bit或32bit。32bit表示直接转换模型，8bit表示模型进行量化。● 量化批大小：以数值形式输入量化批大小。必须为正整数。

7 部署模型

7.1 模型部署简介

在完成训练作业并生成模型后，可在“部署上线”页面对模型进行部署，您也可以将从OBS导入的模型进行部署。ModelArts当前支持如下几种部署类型：

- **在线服务**
将模型部署为一个Web Service，并且提供在线的测试UI与监控能力。
- **批量服务**
批量服务可对批量数据进行推理，完成数据处理后自动停止。
- **边缘服务**
通过华为云智能边缘平台，在边缘节点将模型部署为一个Web Service。

7.2 在线服务

7.2.1 部署为在线服务

模型准备完成后，您可以将模型部署为在线服务，对在线服务进行预测和调用。



用户最多可创建20个在线服务。

前提条件

- 数据已完成准备：已在ModelArts中创建状态“正常”可用的模型。
- 由于在线运行需消耗资源，确保账户未欠费。

操作步骤

1. 登录ModelArts管理控制台，在左侧导航栏中选择“部署上线”，默认进入“在线服务”列表。
2. 在在线服务列表中，单击左上角“部署”，进入“部署”页面。
3. 在“部署”页面，填写在线服务相关参数，然后单击“下一步”。

- a. 填写基本信息，详细参数说明请参见[表7-1](#)。

表 7-1 基本信息参数说明

参数名称	说明
“计费模式”	当前仅支持“按需计费”，不支持修改。
“名称”	在线服务的名称，请按照界面提示规则填写。
“是否自动停止”	启用该参数并设置时间后，服务将在指定时间后自动停止。如果不启用此参数，在线服务将一直运行，同时一直收费，自动停止功能可以帮您避免产生不必要的费用。 目前支持设置为“1小时后”、“2小时后”、“6小时后”、“自定义”。选择“自定义”的模式，可在右侧时间框中选择自动停止时间，指定时间时，必须晚于当前时间。
“描述”	在线服务的简要说明。

图 7-1 部署在线服务基本信息

The screenshot shows the configuration interface for basic information. Key fields include:

- 计费模式:** 按需计费 (Pay-as-you-go) (Selected)
- 名称:** service-99c3
- 是否自动停止:** 启用 (Selected)
- 描述:** (Empty)

A note below the auto-stop toggle states: "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".

- b. 填写资源池和模型配置等关键信息，详情请参见[表7-2](#)。

表 7-2 参数说明

参数名称	子参数	说明
“资源池”	“公共资源池”	公共资源池有CPU或GPU两种规格，不同规格的资源池，其收费标准不同，详情请参见 价格详情说明 。当前仅支持按需付费模式。
“资源池”	“专属资源池”	创建专属资源池请参见 购买专属资源池 。您可以在资源池规格中选择对应的规格进行使用。

参数名称	子参数	说明
“选择模型及配置”	“模型列表”	系统自动关联模型管理中可用的模型列表，选择状态“正常”的模型及版本。
	“分流”	设置当前实例节点的流量占比。 如您仅部署一个版本模型，请设置为100%。如您添加多个版本进行灰度发布，多个版本分流之和设置为100%。
	“计算节点规格”	当选择“公共资源池”时，支持“CPU 2核 8GiB”和“CPU: 2核 8GiB GPU: 1*P4”两种规格。 说明 <ul style="list-style-type: none">● 如选择的是自学习模型及版本，则此处界面可选“自学习规格（CPU）”、“自学习规格（GPU）”。● 规格“CPU: 2核 8GiB GPU: 1*P4”需要提工单申请。
	“计算节点个数”	设置当前版本模型的实例个数。如果节点个数设置为1，表示后台的计算模式是单机模式；如果节点个数设置大于1，表示后台的计算模式为分布式的。请根据实际编码情况选择计算模式。
	“环境变量”	设置环境变量，注入环境变量到容器实例。
	“添加模型版本进行灰度发布”	ModelArts提供多版本支持和灵活的流量策略，您可以通过使用灰度发布，实现模型版本的平滑过渡升级。 说明 如果您选择的模型当前只有一个版本，则界面上不会出现添加模型版本进行灰度发布。

图 7-2 设置模型配置相关信息



- c. 完成参数填写后，单击“下一步”。
4. 在“规格确认”页面，确认填写信息无误后，单击“立即创建”，完成在线服务的部署。部署服务一般需要运行一段时间，根据您选择的数据量和资源不同，训练时间将耗时几分钟到几十分钟不等。

说明

在线服务部署完成后，将立即启动，运行过程中将按照您选择的资源按需计费。

您可以前往在线服务列表，查看在线的基本情况。在在线服务列表中，刚部署的服务“状态”为“部署中”，当在线服务的“状态”变为“运行中”时，表示服务部署完成。

7.2.2 查看服务详情

当模型部署为在线服务成功后，，您可以进入“在线服务”页面，来查看服务详情。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线>在线服务”，进入“在线服务”管理页面。
2. 单击目标服务名称，进入服务详情页面。
 - 您可以查看服务的“名称”、“状态”、“服务ID”、“来源”、“调用失败次数/总次数”、“网络配置”和“描述”。
 - 您也可以通过单击描述右侧的，对描述信息进行编辑。

图 7-3 在线服务详情



3. 您可以在如下页面查看服务相关参数信息，包括“调用指南”、“预测”、“配置更新记录”、“监控信息”、“事件”、“日志”、“共享”、“溯源图”等。

表 7-3 在线服务详情

参数	说明
调用指南	展示API接口地址、模型信息、输入参数、输出参数。您可以通过  复制API接口地址，调用服务。
预测	对在线服务进行预测测试。具体操作请参见 测试服务 。
配置更新记录	展示“当前配置”详情和“历史更新记录”。 <ul style="list-style-type: none">● “当前配置”：模型名称、版本、状态、分流、计算节点规格和计算节点个数。● “历史更新记录”：展示历史模型相关信息。
监控信息	展示当前模型的“资源统计信息”和“模型调用次数统计”。 <ul style="list-style-type: none">● “资源统计信息”：包括CPU、内存、GPU的可用和已用信息。● “模型调用次数统计”：当前模型的调用次数，从模型状态为“已就绪”后开始统计。
事件	展示当前服务使用过程中的关键操作，比如服务部署进度、部署异常的详细原因、服务被启动、停止、更新的时间点等。

参数	说明
日志	展示当前服务下每个模型的日志信息。包含最近5分钟、最近30分钟、最近1小时和自定义时间段。 ● 自定义时间段您可以选择开始时间和结束时间。
共享	展示当前服务共享信息，包括哪些用户订阅了当前服务，以及对当前服务的调用次数等。
溯源图	展示当前服务与数据、训练及模型间的溯源图。 在“溯源图”页面中，您可以查看当前在线服务与模型之间的溯源图。 在溯源图区域，选择任意一个元素，在界面右侧将展示此元素的详细信息。

7.2.3 测试服务

模型部署为在线服务成功后，您可以在“预测”页签进行代码调试或添加图片测试。测试服务包括如下两种方式：

1. **代码预测**：如当前服务是数值类预测，可以在“预测”页签输入代码进行服务预测。
2. **图片预测**：如当前服务是图片识别类预测，可以在“预测”页签添加图片进行服务预测。

代码预测

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线>在线服务”，进入“在线服务”管理页面。
2. 单击目标服务名称，进入服务详情页面。在“预测”页签的预测代码下，输入预测代码，然后单击“预测”即可进行服务的预测，如图7-4所示，attr_7为目标列attr_7的预测结果。

图 7-4 预测代码

The screenshot shows the ModelArts management console interface. At the top, there is a navigation bar with tabs: 调用指南, 配置更新记录, 预测, 监控信息, 日志, 共享, 溯源图. The '预测' tab is currently selected. Below the navigation bar, there is a search bar labeled '请输入路径： /'. On the left, there is a section titled '预测代码' containing the following Python code:

```
6 "req_data": [
7 {
8 "150": 5,
9 "4": 3.3,
10 "setosa": 1.4,
11 "versicolor": 0.2
12 },
13 {
14
15 "150": 5,
16 "4": 2,
17 "setosa": 3.5,
18 "versicolor": 1
19 },
20 {
21 "150": 6,
22 "4": 2.2,
23 "setosa": 5,
24 "versicolor": 1.5
25 }
```

Below the code input area is a red '预测' (Predict) button. To the right, under the heading '返回结果' (Return Result), the JSON response is displayed:

```
3     "uuid": "2964771c-2bc7-40c8-9c01-fee3e8dcd63"
4   },
5   "data": [
6     "resp_data": [
7       {
8         "4": 3.3,
9         "150": 5,
10        "setosa": 1.4,
11        "versicolor": 0.2,
12        "predictresult": 0
13      },
14      {
15        "4": 2,
16        "150": 5,
17        "setosa": 3.5,
18        "versicolor": 1,
19        "predictresult": 0
20      },
21      {
22        "4": 2.2,
```



输入数据中attr_7的值可任意填写，或为空，不会影响预测结果。

图片预测

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线>在线服务”，进入“在线服务”管理页面。
2. 单击目标服务名称，进入服务详情页面。在“预测”页签，单击图片选择按钮 \cdots ，然后选择测试图片。图片上传成功后，单击“预测”即可进行服务的测试，如图7-5所示，输出标签名称“yunbao”，以及位置坐标和检测的评分。

图 7-5 图片预测



7.2.4 访问在线服务

若在线服务的状态处于“运行中”，则表示在线服务已部署成功，您可以使用以下两种方式向在线服务发起预测请求。

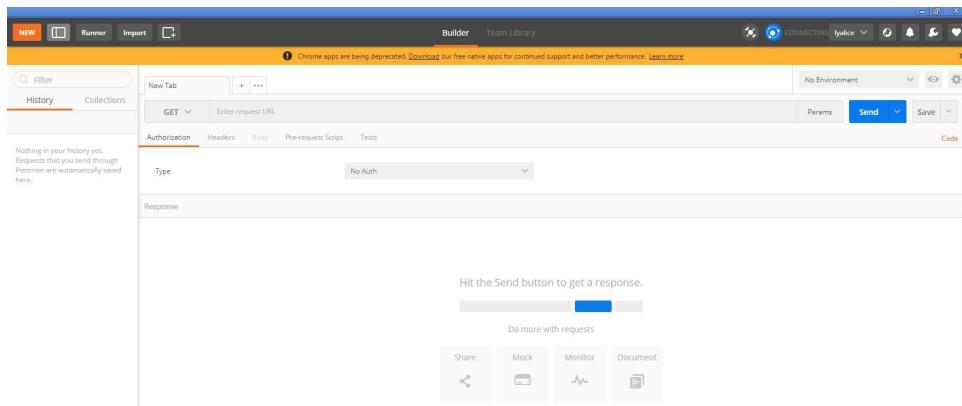
方式一：使用图形界面的软件进行预测（以Postman为例）。

方式二：使用curl命令发送预测请求。

方式一：使用图形界面的软件进行预测（以 Postman 为例）

1. 下载Postman软件并安装，您也可以直接在Chrome浏览器添加Postman扩展程序（也可使用其它支持发送post请求的软件）。
2. 打开Postman，如图7-6所示。

图 7-6 Postman 界面



3. 在Postman界面填写参数，以图像分类举例说明。

- 选择POST任务，将在线服务的调用地址（通过在线服务详情界面-调用指南页签查看）复制到POST后面的方框。Headers页签的Key值填写为“X-Auth-Token”，Value值为您获取到的Token（关于如何获取token，请参考[获取请求认证](#)），如图7-7所示。

说明

您也可以通过AK（Access Key ID）/SK（Secret Access Key）加密调用请求，具体可参见《ModelArts API参考》中的["获取请求认证>AK/SK认证"](#)。

图 7-7 参数填写

The screenshot shows the Postman interface with a POST request. The 'Headers' tab is selected, showing one entry: 'X-Auth-Token' with a value of 'MIITQAVJKoZihvNAQcColTMTCCEy0CAQExDTALBgjhgBZQMEAghGvghGOBgkqhkiG9w...'. Other tabs like 'Params', 'Authorization', 'Body', etc., are visible.

- 在Body页签，根据模型的输入参数不同，可分为2种类型：文件输入、文本输入。

■ 文件输入

选择“form-data”。在“KEY”值填写模型的入参，比如本例中预测图片的参数为“images”。然后在“VALUE”值，选择文件，上传一张待预测图片（当前仅支持单张图片预测），如图7-8所示。

图 7-8 填写 Body

The screenshot shows the Postman interface with a POST request. The 'Body' tab is selected, showing a 'form-data' section. It has one entry: 'images' with a value of 'image_0001.jpg'. There are other options like 'x-www-form-urlencoded', 'raw', and 'binary' available.

■ 文本输入

选择“raw”，选择JSON(application/json)类型，在下方文本框中填写请求体，请求体样例如下：

```
{  
    "meta": {  
        "uuid": "10eb0091-887f-4839-9929-cbc884f1e20e"  
    },  
    "data": {  
        "req_data": [  
            {  
                "sepal_length": 3,  
                "sepal_width": 1,  
                "petal_length": 2.2,  
                "petal_width": 4  
            }  
        ]  
    }  
}
```

其中，“meta”中可携带“uuid”，调用时传入一个“uuid”，返回预测结果时回传此“uuid”用于跟踪请求，如无此需要可不填写meta。

“data”包含了一个“req_data”的数组，可传入单条或多条请求数据，其中每个数据的参数由模型决定，比如本例中的“sepal_length”、“sepal_width”等。

4. 参数填写完成，点击“send”发送请求，结果会在“Response”下的对话框里显示。

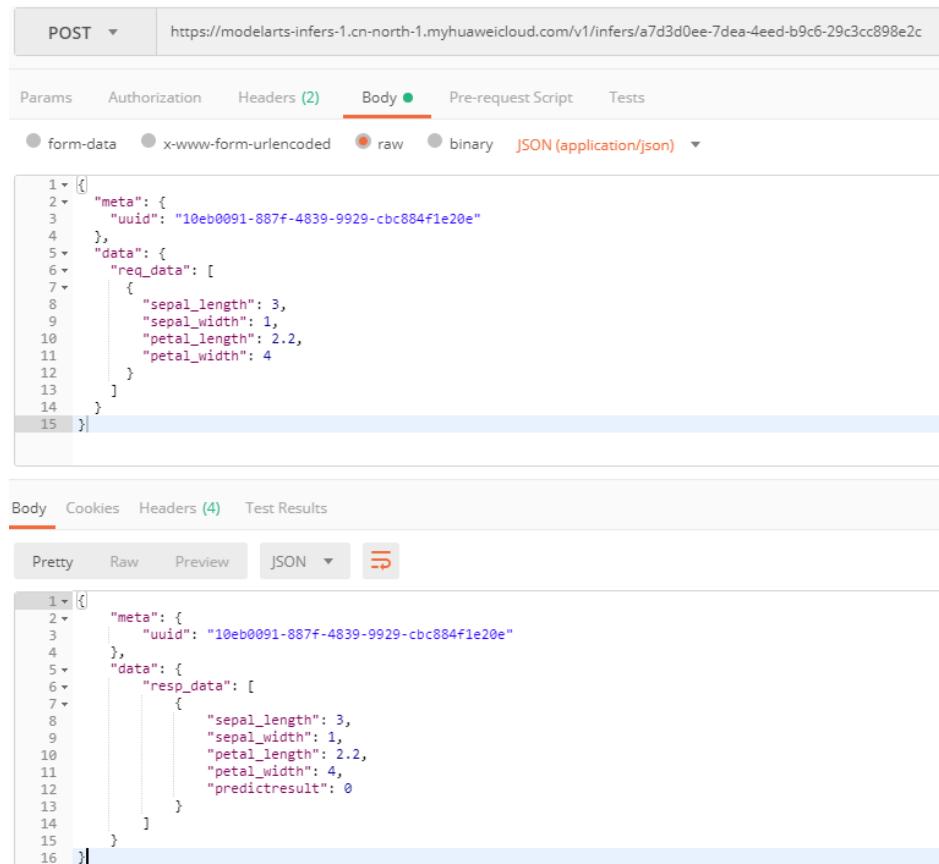
- 文件输入形式的预测结果样例如图7-9所示，返回结果的字段值根据不同模型可能有所不同。
- 文本输入形式的预测结果样例如图7-10所示，请求体包含“meta”及“data”。如输入请求中包含“uuid”，则输出结果中回传此“uuid”。如未输入，则为空。“data”包含了一个“resp_data”的数组，返回单条或多条输入数据的预测结果，其中每个结果的参数由模型决定，比如本例中的“sepal_length”、“predictresult”等。

图 7-9 文件输入预测结果

The screenshot shows a POST request in Postman. The URL is <https://modelarts-infers-1.cn-north-1.myhuaweicloud.com/v1/infers/eb3e0c54-3dfa-4750-af0c-95c45e5d3e83>. The 'Body' tab is selected, showing a file named 'image_0001.jpg' selected under 'form-data'. The 'Pretty' tab displays the JSON response:

```
1  [
2    "confidences": [
3      [
4        0.37127092480659485,
5        0.2595103085041046,
6        0.24806123971939087,
7        0.061120226979255676,
8        0.03235970064997673
9      ]
10     ],
11     "logits": [
12       [
13         1.140504240989685,
14         0.7823686003684998,
15         -1.299513816833496,
16         -0.6635849475860596,
17         -1.455803394317627,
18         0.737247884273529
19       ]
20     ],
21     "labels": [
22       [
23         0,
24         1,
25         5,
26         3,
27         2
28       ]
29     ]
30   ]
```

图 7-10 文本输入预测结果



方式二：使用 curl 命令发送预测请求

使用curl命令发送预测请求的命令格式也分为文件输入、文本输入两类。

1. 文件输入

```
curl -F 'images=@图片路径' -H 'X-Auth-Token:Token值' -X POST 在线服务地址
```

- “-F”是指上传数据的是文件，本例中参数名为“images”，这个名字可以根据具体情况变化，@后面是图片的存储路径。
- “-H”是post命令的headers，Headers的Key值为“X-Auth-Token”，这个名字为固定的，Token值是用户获取到的token值（关于如何获取token，请参考[获取请求认证](#)）。
- “POST”后面跟随的是在线服务的调用地址。

curl命令文件输入样例：

```
curl -F 'images=@/home/data/test.png' -H 'X-Auth-Token:MIISkAY***80T9wHQ==' -X POST https://modelarts-infers-1.cn-north-1.myhuaweicloud.com/v1/infers/eb3e0c54-3dfa-4750-af0c-95c45e5d3e83
```

2. 文本输入

```
curl -d '{"data":{"req_data":[{"sepal_length":3,"sepal_width":1,"petal_length":2.2,"petal_width":4}]}' -H 'X-Auth-Token:MIISkAY***80T9wHQ==' -H 'Content-type: application/json' -X POST https://modelarts-infers-1.cn-north-1.myhuaweicloud.com/v1/infers/eb3e0c54-3dfa-4750-af0c-95c45e5d3e83
```

“-d”是Body体的文本内容。

7.2.5 发布至市场

ModelArts提供了“AI市场”功能，方便将个人的模型、API、数据集等共享给所有ModelArts用户，您也可以从“AI市场”获取他人共享的内容，快速完成构建。在您完成模型的部署之后，您可以将部署的在线服务API发布至“AI市场”，进行知识共享。您也可以在ModelArts控制台“AI市场”中查看其它用户的发布给您的服务API以及您的发布。

前提条件

已在ModelArts中导入模型。且至少存在一个版本。

操作步骤

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线>在线服务”，进入“在线服务”管理页面。
2. 单击目标在线服务操作列的“更多>市场发布”。
3. 在弹出的对话框中填写参数，参数填写如表7-4所示。

表 7-4 参数说明

参数名称	说明
发布者	市场中显示的发布者名称，发布后将不能修改。
名称	市场中显示的模型名称。
描述	对发布API的简要描述，建议从使用场景、使用方法、训练数据集三个方面描述您的API。
模型画像	设置模型画像后，可在市场中显示模型的标签，便于分类显示和快速查找到API。 建议从每个属性的菜单中选择至多三个最合适的标签用于描述您的API，如没有合适的标签，也可留空不选。可以从行业、数据、场景、主题、模型、框架引擎几个类型中选择标签。
封面图	封面图可帮助其他用户直观感受您的模型的用途，您可以从“OBS”中选取，有可以从“本地上传”，如果您已共享过模型也可以使用上一次的封面图。 GIF动态图会是非常好的选择，支持jpg、png、gif、bmp格式的图像，最佳宽高比是“5:3”。
发布到	可共享到“AI市场”，也可共享给“个人”。 <ul style="list-style-type: none">● “AI市场”：共享给所有用户。共享到AI市场需要审核，可在共享对话框中查看审批状态。● “个人”：共享给指定用户。 说明<ul style="list-style-type: none">● 账号ID可从“我的凭证”中获取。● 输入多个用户ID时，用户ID之间请用“，”隔开，且不允许出现特殊字符及空格。

- 单击“确定”，完成API的发布操作。

您可以登录ModelArts管理控制台，“AI市场>我的发布”页面查看您的发布，查看您的发布操作请参见[我发布的](#)。

7.3 批量服务

7.3.1 部署为批量服务

模型准备完成后，您可以将模型部署为批量服务。在部署上线“批量服务”界面，列举了用户所创建的批量服务。您可以在右上方搜索框中输入服务名称，单击进行查询。

前提条件

- 数据已完成准备：已在ModelArts中创建状态“正常”可用的模型。
- 准备好需要批量处理的数据，并上传至OBS目录。
- 已在OBS创建至少1个空的文件夹，用于存储训练输出的内容。

背景信息

- 批量服务目前还处于限时免费阶段，运行中的批量服务，并不会产生费用。
- 用户最多可创建1000个批量服务。

操作步骤

- 登录ModelArts管理控制台，在左侧导航栏中选择“部署上线”，默认进入“批量服务”列表。
- 在批量服务列表中，单击左上角“部署”，进入“部署”页面。
- 在部署页面，填写批量服务相关参数，然后单击下一步。
 - 填写基本信息。基本信息包含“名称”、“描述”。其中“名称”默认生成。例如：service-bc0d，您也可以根据实际情况填写“名称”和“描述”信息。
 - 填写服务参数。包含资源池、模型配置等关键信息，详情请参见[表7-5](#)。

表 7-5 参数说明

参数名称	说明
选择模型及版本	选择状态“正常”的模型及版本。
输入数据目录位置	选择输入数据的OBS路径，即您上传数据的OBS目录。只能选择文件夹或“.manifest”文件。 “.manifest”文件规范请参见 Manifest文件规范 。
请求路径	批量服务中调用模型的接口URL。

参数名称	说明
映射关系	填写每个参数对应到csv单行数据的字段索引，索引index从0开始计数。 根据model文件自动生成映射关系。 当model文件中包含“file=images”、“data=json”任一种信息会出现映射关系详情。
输出数据目录位置	选择批量预测结果的保存位置，可以选择您创建的空文件夹。
计算节点规格	“CPU 2核 8GiB” 和 “CPU: 2核 8GiB GPU: 1*P4”两种规格。 说明 <ul style="list-style-type: none">如选择的是自动学习模型及版本，则此界面可选“自动学习规格（CPU）”、“自动学习规格（GPU）”。规格“CPU: 2核 8GiB GPU: 1*P4”需要提工单申请。
计算节点个数	设置当前版本模型的实例个数。如果节点个数设置为1，表示后台的计算模式是单机模式；如果节点个数设置大于1，表示后台的计算模式为分布式的。请根据实际编码情况选择计算模式。
环境变量	设置环境变量，注入环境变量到容器实例。

4. 完成参数填写后，单击“立即创建”，完成批量服务的部署。部署服务一般需要运行一段时间，根据您选择的数据量和资源不同，训练时间将耗时几分钟到几十分钟不等。

说明

批量服务部署完成后，将立即启动，运行过程中将按照您选择的资源按需计费。
您可以前往批量服务列表，查看批量服务的基本情况。在批量服务列表中，刚部署的服务“状态”为“部署中”，当批量服务的“状态”变为“运行完成”时，表示服务部署完成。

Manifest 文件规范

推理平台批量服务支持使用manifest文件，manifest文件可用于描述数据的输入输出。

输入manifest文件样例

- 文件名：“test.manifest”
- 文件内容：

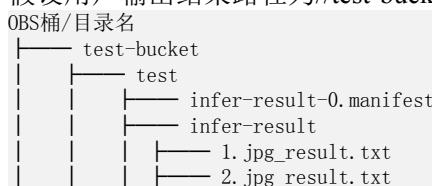
```
{"source": "s3://obs-data-bucket/test/data/1.jpg"}  
{"source": "https://infers-data.obs.cn-north-1.myhwclouds.com:443/xgboosterdata/data.csv"}  
AccessKeyId=2Q0V0TQ461N26DDL18RB&Expires=1550611914&Signature=wZBtZj5QZrReDhz1uDzwve8GpY  
%3D&x-obs-security-token=gQpbz3V0aGNoaW5hixvY8V9a1SnsxmGoHYmB1SArYMyqnQT-  
ZaMSxHv168kKLay5feYvLDMNZWxzhBZ6Q-3HcoZMh9gISwQ0VBwm4ZytB_m8sg1fL6isU7T3CnoL9jmvdGgT9VBC7dC1E  
yfSJrUcqfB_N0ykCsfrAITt_IQYZFDu_HyqVk-  
GunUcTVdDfW1CV3TrYcpmnZjliAnYu089kAwCYGeRzsCsC0ePu4PHMsBvYV9gWmN9AUZIDn1sfRL4voBpwQnp6tnAgHW  
49y5a6hP2hCAoQ-95SpUriJ434QlymoeKfTHVMKOeZxZea-  
Jx0vevOCGI5CcGehEJaz48sgH81UiHz121zocNB_hpPfus2jY6KPglEJxMv6Kwmro-ZBXWuSJUDOnSYXI-3ciYjg9-  
h1b8W3sW1mOTFCWNGoWsd74it7l_5-7UUhoIeyPBy0_ReWkcur2F0JsuMpGRaPyg1ZxXm_jfdLFXobYtzZhbui4yxWXga  
6oxT0kfewykTOYH0NPoPr5MYGYweOXxsFs3d5w2rd0y7p0QYhTzIkk5C1z7F1WNapFISL7zdhs18RfchTqESq94Kgke  
qatSF_i1vnYMW2r8P8x2k_eb6NJ7U_q5ztMb09oWEcfrr0D2f7n7B1_nb2HIB_H9tjzKvqwnngaimYhBbMRPfibvttW86Gi  
wVP8vrC27F0n39Be9z2hSfJ_8pHej0yM1yNqZ481FQ5vWT_vFV3JHM-7I1ZB0_hIdaHfItm-  
J69cTfHSE0zt7DGaMIES1o7U3w%3D%3D"}
```

- 文件要求：
 - a. 文件名后缀需为“.manifest”；
 - b. 文件内容是多行JSON，每行JSON描述一个输入数据，需精确到文件，不能是文件夹；
 - c. JSON内容需定义一个source字段，字段值是OBS的文件地址，有2种表达形式：
 - i. “s3://{{桶名}}/{{对象名}}”，适用于访问自己名下的OBS数据；
 - ii. OBS生成的分享链接，包含签名信息。适用于访问其他人的OBS数据。

输出manifest文件样例

使用manifest文件输入，输出结果目录也会有一个manifest文件。

- 假设用户输出结果路径为//test-bucket/test/，则结果存放位置如下：



- infer-result-0.manifest文件内容：

```
{"source": "s3://obs-data-bucket/test/data/1.jpg", "inference-loc": "s3://test-bucket/test/infer-result/1.jpg_result.txt"} {"source": "https://infers-data.obs.cn-north-1.myhwclouds.com:443/xgboosterdata/2.jpg?AccessKeyId=2Q0VOTQ461N26DDL18RB&Expires=1550611914&Signature=wZBttZj5QzrReDhzluDzwve8GpY%3D&x-obs-security-token=gQpbz3V0aGNoaW5hixvY8V9a1SnsxmGoHYmB1SArYMyqnQT-ZaMSxHv168kKLay5feYvLDMNZWxzhBZ6Q-3HcoZMh9gISwQ0Vbw4ZytB_m8sg1fL6isU7T3CnoL9jmvdGgT9VBC7dC1EyfSJrUcqfB_N0ykCsfrA1T_IQYZFDu_HyqVkJunUcTvDfW1CV3TrYcpmznZj1iAnYu089kAwCYGeRzsCsCoEPu4PHMsBvY9gWmN9AUZIDn1sfRL4voBpwQmp6tnAgHW49y5a6hP2hCaQ-95SpUrj434Q1ymoeKfTHVMK0eZxZea-Jx0vev0CGI5CcGehEJaz48sgH81UiHzl21zocNB_hpPfus2jY6KPglEJxMv6Kwmro-ZBXWuSJUDOnSYXI-3ciYjg9-h10b8W3sW1mOTFCWNGoWsd74it71_5-7UUh0IeyPBBy0_REwkur2F0JsuMpG1RaPyg1ZxXm_jfdLFXobYtzZhbui4yWXga6oxT0kfewykTOYH0NPoPRt5MYGywe0XXxFs3d5w2rd0y7p0QYhyTzIk5Clz7F1WNapF1SL7zdhsl8RfcTqESq94KgkeqatSF_iIvnYMW2r8P8x2k_eb6NJ7U_q5ztMb09oWEcfr0D2f7n7B1_nb2HIB_H9tjzKvqwnqaimYhBbMRPfibvttW86GiwWP8vrC27F0n39Be9z2hSfJ_8pHej0yMlyNqZ481FQ5vWT_vFV3JHM-7I1ZB0_hIdahfItm-J69cTfHSE0zt7DGaMIES1o7U3w%3D%3D", "inference-loc": "s3://test-bucket/test/infer-result/2.jpg_result.txt"}
```

- 文件格式：

- a. 文件名为“infer-result-{{index}}.manifest”，index为实例序号，批量服务运行多少个实例就会产生多少个manifest文件；
- b. manifest同一目录下会创建infer-result目录存放结果；
- c. 文件内容是多行JSON，每行JSON描述一个输入数据的对应输出结果；
- d. JSON内容包含2个字段，source、inference-loc：
 - i. source：输入数据描述，与输入的manifest一致；
 - ii. inference-loc：输出结果路径，格式为“s3://{{桶名}}/{{对象名}}”。

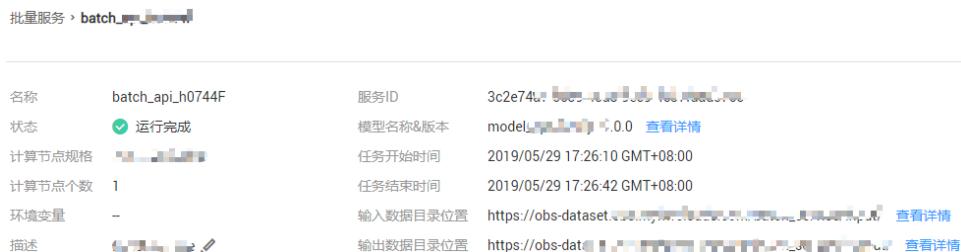
7.3.2 查看批量服务预测结果

当您在部署批量服务时，会选择输出数据目录位置，您可以查看“运行完成”状态的批量服务运行结果。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线>批量服务”，进入“批量服务”管理页面。
2. 单击状态为“运行完成”的目标服务名称，进入服务详情页面。

- 您可以查看服务的“名称”、“状态”、“服务ID”、“输入数据目录位置”、“输出数据目录位置”、“网络配置”和“描述”。
- 您也可以通过单击描述右侧的，对描述信息进行编辑。

图 7-11 批量服务详情



批量服务：batch_api_h0744F	
名称	batch_api_h0744F
状态	 运行完成
计算节点规格	
计算节点个数	1
环境变量	-
描述	
服务ID	3c2e74...
模型名称&版本	model_... v.0.0 查看详情
任务开始时间	2019/05/29 17:26:10 GMT+08:00
任务结束时间	2019/05/29 17:26:42 GMT+08:00
输入数据目录位置	https://obs-dataset... 查看详情
输出数据目录位置	https://obs-dataset... 查看详情

3. 单击“输出数据目录位置”后的“查看详情”，可以获取批量服务预测结果。
 - 当输入为图片时，每张图片输出一个结果，输出结果格式为：图片名_result.txt。例如：IMG_20180919_115016.jpg_result.txt。
 - 当输入为音频时，每个音频输出一个结果，输出结果格式为：音频名_result.txt。例如：1-36929-A-47.wav_result.txt。
 - 当输入为表格数据时，输出结果格式为：表格名_result.txt。例如：train.csv_result.txt。

7.4 边缘服务

7.4.1 部署为边缘服务

模型准备完成后，您可以将模型部署为边缘服务。在部署上线“边缘服务”界面，列举了用户所创建的边缘服务。您可以在右上方搜索框中输入服务名称，单击进行查询。边缘服务依赖华为云智能边缘平台（IEF），部署前需要在智能边缘平台上创建边缘节点。

前提条件

- 数据已完成准备：已在ModelArts中创建状态“正常”可用的模型。
- 已在智能边缘平台上创建边缘节点，如果您未创建边缘节点，具体操作请参见[创建边缘节点](#)。
- 由于在线运行需消耗资源，确保账户未欠费。

背景信息

- 边缘服务目前还处于限时免费阶段，运行中的边缘服务，并不会产生费用。
- 用户最多可创建1000个边缘服务。

操作步骤

1. 登录ModelArts管理控制台，在左侧导航栏中选择“部署上线”，默认进入“边缘服务”列表。
2. 在边缘服务列表中，单击左上角“部署”，进入“部署”页面。

3. 在部署页面，填写边缘服务相关参数，然后单击下一步。
 - a. 填写基本信息。基本信息包含“名称”、“描述”。其中“名称”默认生成。例如：service-bc0d，您也可以根据实际情况填写“名称”和“描述”信息。
 - b. 填写服务参数。包含资源池、模型配置等关键信息，详情请参见[表7-6](#)。

表 7-6 参数说明

参数名称	说明
“选择模型及配置”	选择状态“正常”的模型及版本。
“计算节点规格”	支持“CPU: 2核 8GiB”和“CPU: 2核 8GiB GPU: 1*P4”两种规格。 说明 <ul style="list-style-type: none">● 如选择的是自动学习模型及版本，则此界面可选“自动学习规格（CPU）”、“自动学习规格（GPU）”。● 规格“CPU: 2核 8GiB GPU: 1*P4”需要提工单申请。
“环境变量”	设置环境变量，注入环境变量到容器实例。
“选择边缘节点”	边缘节点是您自己的边缘计算设备，用于运行边缘应用，处理您的数据，并安全、便捷地和云端应用进行协同。 单击选择边缘节点“添加”，在弹出的“添加节点”对话框中选择节点。选择您已创建的节点后，单击“确定”。

4. 完成参数填写后，单击“立即创建”，完成边缘服务的部署。部署服务一般需要运行一段时间，根据您选择的数据量和资源不同，部署时间将耗时几分钟到几十分钟不等。
您可以前往边缘服务列表，查看边缘服务的基本情况。在边缘服务列表中，刚部署的服务“状态”为“部署中”，当边缘服务的“状态”变为“运行中”时，表示服务部署完成。

7.4.2 访问边缘服务

访问边缘服务

当边缘服务和边缘节点的状态都处于“运行中”状态，表示边缘服务已在边缘节点成功部署。

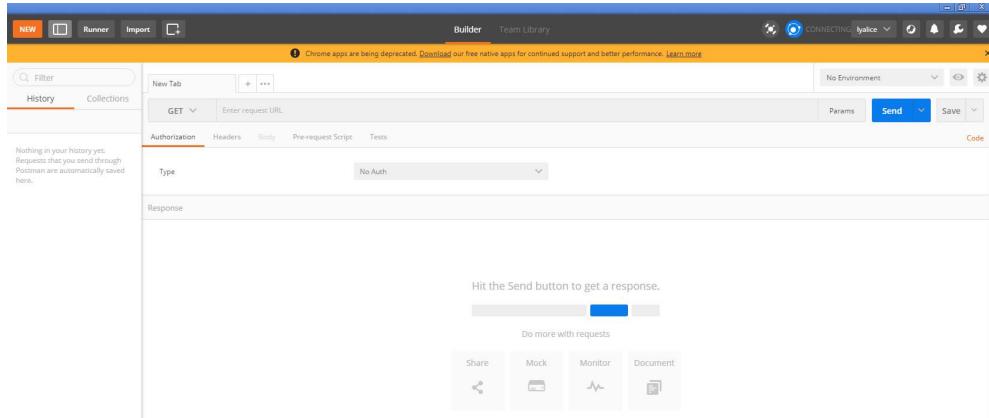
您可以通过以下两种方式，在能够访问到边缘节点的网络环境中，对部署在边缘节点上的边缘服务发起预测请求。

- 方式一：使用图形界面的软件进行预测（以Postman为例）
- 方式二：使用curl命令发送预测请求

方式一：使用图形界面的软件进行预测（以 Postman 为例）

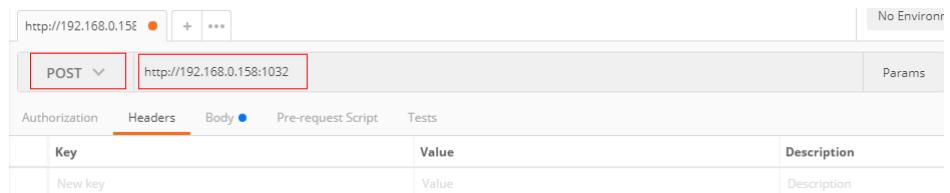
1. 下载Postman软件并安装，您可以直接在Chrome浏览器添加Postman扩展程序（也可使用其它支持发送post请求的软件）。
2. 打开Postman，如图7-12所示。

图 7-12 Postman 界面



3. 在Postman界面填写参数，以图像分类举例说明。
 - 选择POST任务，将某个边缘节点的调用地址（通过边缘服务详情界面-节点信息页签查看）复制到POST后面的方框。

图 7-13 参数填写

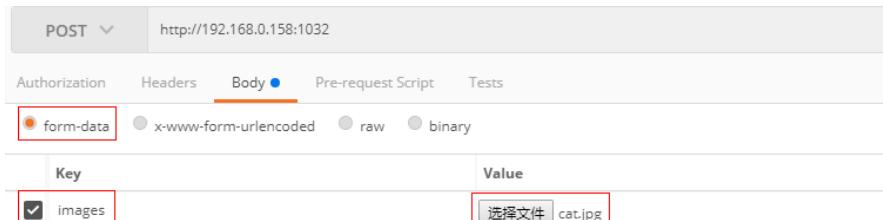


- 在Body页签，根据模型的输入参数不同，可分为2种类型：文件输入、文本输入。

■ 文件输入

选择“form-data”。在“KEY”值填写模型的入参，比如本例中预测图片的参数为“images”。然后在“VALUE”值，选择文件，上传一张待预测图片（当前仅支持单张图片预测）。

图 7-14 填写 Body



■ 文本输入

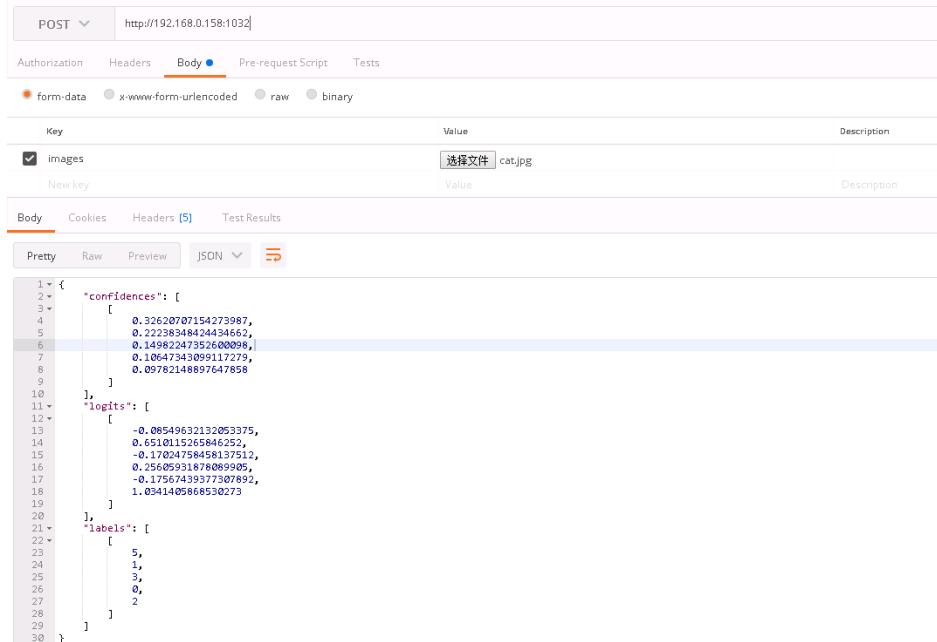
选择“raw”，选择JSON(application/json)类型，在下方文本框中填写请求体，请求体样例如下。

```
{  
    "meta": {  
        "uuid": "10eb0091-887f-4839-9929-cbc884f1e20e"  
    },  
    "data": {  
        "req_data": [  
            {  
                "sepal_length": 3,  
                "sepal_width": 1,  
                "petal_length": 2.2,  
                "petal_width": 4  
            }  
        ]  
    }  
}
```

其中，“meta”中可携带“uuid”，返回预测结果时回传此“uuid”用于跟踪请求，如无此需要可不填写meta。“data”包含了一个“req_data”的数组，可传入单条或多条请求数据，其中每个数据的参数由模型决定，比如本例中的“sepal_length”、“sepal_width”等。

4. 参数填写完成，点击“Send”发送请求，结果会在Response下的对话框里显示。
 - 文件输入形式的预测结果样例如图7-15所示，返回结果的字段值根据不同模型可能有所不同。

图 7-15 文件输入预测结果



- 文本输入形式的预测结果样例如图7-16所示，请求体包含“meta”及“data”。如输入请求中包含“uuid”，则输出结果中回传此“uuid”。如未输入，则为空。“data”包含了一个“req_data”的数组，可传入单条或多条请求数据，其中每个数据的参数由模型决定，比如本例中的“sepal_length”、“sepal_width”等。

图 7-16 文本输入预测结果

The screenshot shows a POST request in Postman. The URL is `http://192.168.0.158:1033`. The request body is a JSON object:

```
1 {  
2   "meta": {  
3     "uuid": "10eb0091-887f-4839-9929-cbc884f1e20e"  
4   },  
5   "data": {  
6     "req_data": [  
7       {  
8         "sepal_length": 3,  
9         "sepal_width": 1,  
10        "petal_length": 2.2,  
11        "petal_width": 4  
12      }  
13    ]  
14  }  
15 }
```

The response body is also a JSON object:

```
1 [ ]  
2 {  
3   "meta": {  
4     "uuid": "10eb0091-887f-4839-9929-cbc884f1e20e"  
5   },  
6   "data": {  
7     "res_data": [  
8       {  
9         "sepal_length": 3,  
10        "sepal_width": 1,  
11        "petal_length": 2.2,  
12        "petal_width": 4,  
13        "predictresult": 0  
14      }  
15    ]  
16  }  
17 }
```

方式二：使用 curl 命令发送预测请求

使用curl命令发送预测请求的命令格式也分为文件输入、文本输入两类

1. 文件输入

```
curl -F 'images=@图片路径' -X POST 边缘节点服务地址
```

- “-F”是指上传数据的是文件，本例中参数名为**images**，这个名字可以根据具体情况变化，@后面是图片的存储路径。
- “POST”后面跟随的是边缘节点的调用地址。

curl命令文件输入预测样例：

```
curl -F 'images=@/home/data/cat.jpg' -X POST http://192.168.0.158:1032
```

预测结果如图7-17所示。

图 7-17 curl 命令文件输入预测结果

```
root@modelarts006:/# curl -F 'images=@/home/data/cat.jpg' -X POST http://192.168.0.158:1032  
{"confidences": [[0.32620707154273887, 0.22238348424434662, 0.14982247352600098, 0.10647343099117279, 0.0978214889764785  
81], "logits": [[-0.08549632132053375, 0.6510115265846252, -0.17024750458137512, 0.25605931878008905, -0.175674393773078  
92, 1.0341405668530273]], "labels": [[5, 1, 3, 0, 2]]}root@modelarts006:/#
```

2. 文本输入

```
curl -d '{  
  "meta": {  
    "uuid": "10eb0091-887f-4839-9929-cbc884f1e20e"  
  },  
  "data": {  
    "req_data": [  
      {  
        "sepal_length": 3,  
        "sepal_width": 1,  
        "petal_length": 2.2,  
        "petal_width": 4  
      }  
    ]  
  }  
}' http://192.168.0.158:1032
```

```
"data": {
  "req_data": [
    {
      "sepal_length": 3,
      "sepal_width": 1,
      "petal_length": 2.2,
      "petal_width": 4
    }
  ]
}
}' -X POST <边缘节点服务地址>
```

- “-d” 是Body体的文本内容，如模型为文本输入，则需要用此参数。

curl命令文本输入预测样例：

```
curl -d '{
  "meta": {
    "uuid": "10eb0091-887f-4839-9929-cbc884f1e20e"
  },
  "data": {
    "req_data": [
      {
        "sepal_length": 3,
        "sepal_width": 1,
        "petal_length": 2.2,
        "petal_width": 4
      }
    ]
  }
}' -X POST http://192.168.0.158:1033
```

预测结果如图7-18所示。

图 7-18 curl 命令文本输入预测结果

```
root@modelarts006:/# curl -X POST \
>   http://192.168.0.158:1033/ \
> -d '{
>   "meta": {
>     "uuid": "10eb0091-887f-4839-9929-cbc884f1e20e"
>   },
>   "data": {
>     "req_data": [
>       {
>         "sepal_length": 3,
>         "sepal_width": 1,
>         "petal_length": 2.2,
>         "petal_width": 4
>       }
>     ]
>   }
> }'
{"meta": {"uuid": "10eb0091-887f-4839-9929-cbc884f1e20e"}, "data": {"req_data": [{"sepal_length": 3, "sepal_width": 1, "petal_length": 2.2, "petal_width": 4, "predictresult": 0}]}}

root@modelarts006:/#
```

7.5 修改服务

对于已部署的服务，您可以修改服务的基本信息以匹配业务变化。您可以通过如下两种方式修改服务的基本信息：

方式一：通过服务管理页面修改服务信息

方式二：通过服务详情页面修改服务信息

前提条件

已存在部署完成的服务。

方式一：通过服务管理页面修改服务信息

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线”，进入目标服务类型管理页面。
2. 在服务列表中，单击目标服务操作列的“修改”，修改服务基本信息，然后单击“确定”完成修改。
 - 在线服务参数说明请参见[部署为在线服务](#)。
 - 批量服务参数说明请参见[部署为批量服务](#)。
 - 边缘服务参数说明请参见[部署为边缘服务](#)。



“部署中”状态的服务无法进行修改。

方式二：通过服务详情页面修改服务信息

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线”，进入目标服务类型管理页面。
2. 单击目标服务名称，进入服务详情页面。
3. 您可以通过单击页面右上角“修改”，修改服务基本信息，然后单击“确定”完成修改。
 - 在线服务参数说明请参见[部署为在线服务](#)。
 - 批量服务参数说明请参见[部署为批量服务](#)。
 - 边缘服务参数说明请参见[部署为边缘服务](#)。

图 7-19 服务操作



7.6 启动或停止服务

启动服务

您可以对处于“运行完成”、“异常”和“停止”状态的服务进行“启动”操作，“部署中”状态的服务无法启动。启动服务，ModelArts将开始计费。您可以通过如下两种方式启动服务：

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线”，进入目标服务类型管理页面。您可以单击“操作”列的“启动”，启动服务。
2. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线”，进入目标服务类型管理页面。单击目标服务名称，进入服务详情页面。您可以单击页面右上角“启动”，启动服务。

停止服务

您可以对处于“运行中”和“告警”状态的服务进行“停止”操作，“部署中”状态的服务无法停止。停止服务，ModelArts将停止计费。您可以通过如下两种方式停止服务：

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线”，进入目标服务类型管理页面。您可以单击“操作”列的“停止”，停止服务。
2. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线”，进入目标服务类型管理页面。单击目标服务名称，进入服务详情页面。您可以单击页面右上角“停止”，停止正在运行中服务。

7.7 删除服务

如果服务不再使用，您可以删除服务释放资源。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线”，进入目标服务类型管理页面。
 - a. 在线服务，您可以单击“操作”列的“更多>删除”，删除服务。
 - b. 批量服务和边缘服务，您可以单击“操作”列的“删除”，删除服务。
2. 登录ModelArts管理控制台，在左侧菜单栏中选择“部署上线”，进入目标服务类型管理页面。单击目标服务名称，进入服务详情页面。您可以单击页面右上角“停止”，停止正在运行中服务。



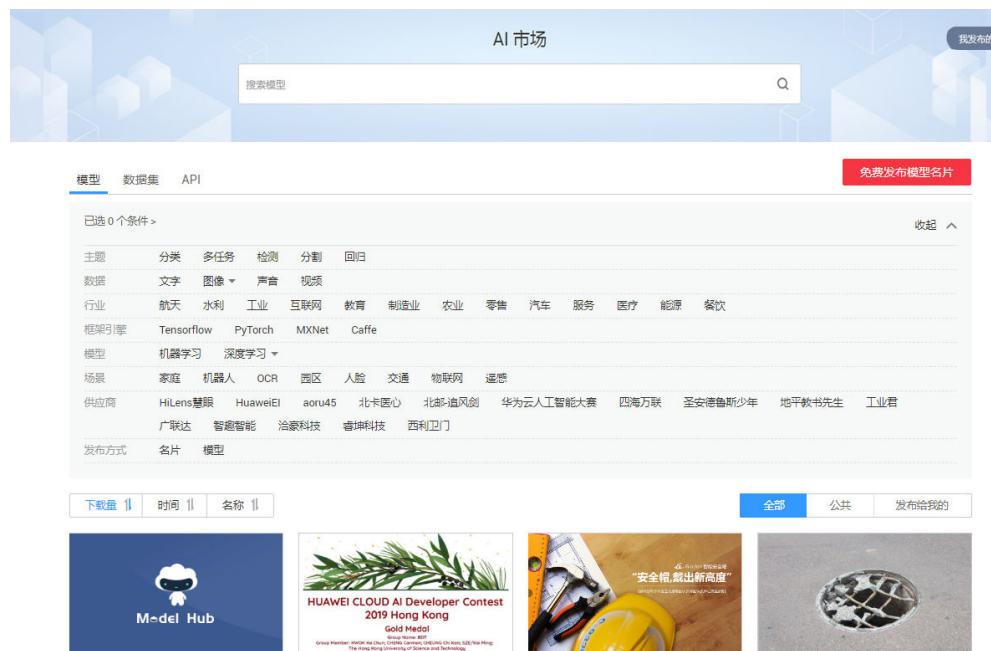
说明

删除操作无法恢复，请谨慎操作。

8 AI 市场

ModelArts的AI市场提供了常用数据集，列举了其他用户共享的模型、API，您可以使用他人分享的信息快速构建模型。同时，您也可以将自己的API或模型发布至AI市场，共享知识。

图 8-1 AI 市场



模型

在ModelArts管理控制台，单击“AI市场”，在“模型”页签中，您可以选择需要的模型。模型页签您可以执行如下操作。

- 搜索模型：您可以在搜索框中输入模型名称或模型类型，单击 进行查询，下方区域会展示和搜索词相关的模型。您也可以通过下方的筛选条件进行模型搜索。
- 查看模型：您可以单击模型名称进入模型详情页面，该页面展示模型的基本信息、使用说明等信息。

- 发布模型名片：单击页面右上方“免费发布模型名片”发布模型信息用于企业业务推广。如需发布可推理模型，请从ModelArts“模型管理”页面进行导入模型后，再发布。发布方式为“名片”的模型信息不包含具体模型。
- 导入至我的模型：单击模型名称进入模型详情页面，单击“导入至我的模型”，在“导入至我的模型”对话框中填写名称、版本和描述，将当前模型导入至“模型管理”。

 **说明**

模型名称支持1-48位可见字符，只能以英大小写字母或者中文字符开头，可以包含字母、中文、数字、中划线、下划线。

版本不能为空，模型版本示例0.0.1。

数据集

在ModelArts管理控制台，单击“AI市场”，在“数据集”页签中，列举了ModelArts预置的所有数据集。数据集页签您可以执行如下操作。

- 搜索数据集：您可以在搜索框中输入数据集名称或数据集类型，单击进行查询，下方区域会展示和搜索词相关的数据集。
- 查看数据集：您可以单击数据集名称进入数据集详情页面，该页面展示数据集的数据样例、评论等信息。
- 导入至我的数据集：单击数据集名称进入数据集详情页面，单击“导入至我的数据集”，在“导入至我的数据集”对话框中填写名称、存储路径和描述，将当前数据集导入至OBS桶中，存储路径选择OBS桶。当前数据集将导入至“数据管理>数据集”。

API

在ModelArts管理控制台，单击“AI市场”，在“API”页签中，您可以选择需要的API。API页签您可以执行如下操作。

- 搜索API：您可以在搜索框中输入API名称或API类型，单击进行查询，下方区域会展示和搜索词相关的API。
- 查看API：您可以单击API名称进入API详情页面，该页面展示API的详细信息。
- 订阅API：单击API名称进入API详情页面，单击“订阅”，对话框中填写名称，将当前API订阅至您的在线服务。

我发布的

在ModelArts管理控制台，进入“AI市场”，单击右上角“我发布的”，可进入“我发布的”页面，列举了您发布的模型和API。如图8-2所示。在“我发布的”页面您可以查看已发布数据的分类、名称、当前状态、评分、下载次数，还可以对已发布数据进行操作和搜索。

图 8-2 我发布的

The screenshot shows a table listing 10 published items (models and APIs) from a total of 77. The columns include Category, Name, Current Status, Score, Download次数 (Downloads), Submission Time, and Operation (Cancel Release). All items are marked as '已发布' (Published).

分类	名称	当前状态	评分	下载次数	申请时间	操作
模型	model_model	已发布	-	1	-	取消发布
模型	xtjaddadl	已发布	-	1	-	取消发布
模型	resnet_v1_50_0117	已发布	-	1	-	取消发布
API	imgClass_v022_service	已发布	-	1	-	取消发布
模型	getmix	已发布	-	1	-	取消发布
API	resnet_v1_50_0119	已发布	-	2	-	取消发布
模型	FIRR_v2_50_0119	已发布	-	1	-	取消发布
API	gettwtest	已发布	-	1	-	取消发布
模型	model	已发布	-	3	-	取消发布
模型	getxG	已发布	-	0	-	取消发布

9 资源池

ModelArts 资源池说明

在使用ModelArts进行AI全流程开发时，您可以选择使用两种不同的资源池训练和部署模型。

- **公共资源池：**公共资源池提供公共的大规模计算集群，根据用户作业参数按需分配使用，资源按作业隔离。按资源规格、使用时长及实例数计费，不区分任务（训练作业、部署、开发）。公共资源池是ModelArts默认提供，不需另行创建或配置，您可以直接在AI开发过程中，直接选择公共资源池进行使用。
- **专属资源池：**提供独享的计算资源，可用于Notebook、训练作业、部署模型等。专属资源池不与其他用户共享，更加高效。

在使用专属资源池之前，您需要先购买一个专属资源池，然后在AI开发过程中选择此专属资源池。专属资源池的详细介绍和操作请参见[专属资源池介绍、购买专属资源池、扩缩容专属资源池、删除专属资源池](#)。



说明

专属资源池涉及昂贵的计算资源，目前处于邀测中，如有需要，请填写[工单](#)申请开通。

专属资源池介绍

- 专属资源池可以在如下作业和任务中使用：Notebook、训练作业、TensorBoard、部署上线（含在线服务、离线服务和边缘服务）。
- 专属资源池分为“开发环境/训练专用”和“部署上线专用”两种类型。“开发环境/训练专用”类型的专属资源池只能用于Notebook、训练作业、TensorBoard等功能，“部署上线专用”类型的专属资源池只能用于模型的部署上线。
- 只有处于“运行中”状态的专属资源池才是可用的。如果专属资源池状态为“不可用”或“异常”，请排除故障后再使用。
- 创建专属资源池后，就会基于选择的规格开始计费。
- 专属资源池的收费支持“按需计费”和“包年包月”两种，具体收费详情请参见[价格详情说明](#)。

购买专属资源池

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“专属资源池”。
2. 在专属资源池管理页面，您可以选择通过“开发环境/训练专用”和“部署上线专用”页签选择两种不同类型的专属资源池。

3. 单击左上角“创建”，进入购买专属资源池界面。
4. 在“购买专属资源池”界面填写参数，参数填写请参见[表9-1](#)和[表9-2](#)。

表 9-1 “开发环境/训练专用”专属资源池的参数说明

参数名称	说明
资源类型	系统默认为“开发环境/训练专用”，不可修改。
计费模式	选择计费模式，“包年/包月”或“按需计费”。
名称	专属资源池的名称。 名称由小写字母、数字、中划线和下划线组成，只能以小写字母开头，且不能以中划线或下划线结尾。
描述	专属资源池的简要描述。
节点数	选择专属资源池的节点数，选择的节点数越多，计算性能越强，同时费用越高。
节点规格	当前仅支持“modelarts.vm.gpu.p100 56核 512GiB 1*P100”节点规格。
购买时长	选择购买时长。只有选择“包年/包月”计费模式时才需填写。 最少为1个月，最长为1年。其中，ModelArts推出套餐包优惠，购买10个月赠送2个月，直接选择1年的购买时长即可完成买10个月送2个月的套餐包购买。

表 9-2 “部署上线专用”专属资源池的参数说明

参数名称	说明
资源类型	系统默认为“部署上线专用”，不可修改。
计费模式	“部署上线专用”类型的专属资源池仅支持“按需计费”。
名称	专属资源池的名称。 名称由小写字母、数字、中划线和下划线组成，只能以小写字母开头，且不能以中划线或下划线结尾。
描述	专属资源池的简要描述。
自定义网络配置	启用自定义配置，则服务实例运行在指定的网络中，可以与该网络中的其它云服务资源实例互通；不启用自定义配置，ModelArts会为每个用户分配一个专属的网络，用户之间隔离。 如果启用自定义网络配置，请设置对应的“虚拟私有云”、“子网”和“安全组”。如果没有可用网络，请前往虚拟私有云服务创建。

参数名称	说明
节点数	选择专属资源池的节点数，选择的节点数越多，计算性能越强，同时费用越高。
节点规格	当前支持“modelarts.vm.cpu.8ud 8核 32GiB”和“modelarts.vm.gpu.p4u8 8核 32GiB 1*P4”两种节点规格，请根据需求选择。

5. 单击“下一步”，进入规格确认。
6. 规格确认无误后，单击“去支付”，然后在支付页面完成付款。
付款成功后即完成专属资源池的购买，您可以在专属资源池列表查看已购买的资源池。当专属资源池创建成功后，其状态将变为“运行中”。

扩缩容专属资源池

当专属资源池使用一段时间后，由于AI开发业务的变化，您可以通过扩容或缩容操作，增加或减少节点数量。

“包年/包月”的专属资源池不支持扩缩容。如果购买的是“按需计费”的专属资源池，那么计费会按照修改后的节点数量进行收费。

扩缩容的操作步骤如下所示：

1. 进入专属资源池管理页面，在专属资源池所在行，单击操作列“扩缩容”。
2. 在扩缩容页面，增加或减少节点数量。增加节点数量表示扩容，减少节点数量表示缩容。请根据本身业务诉求进行调整。
 - 扩容时，请务必选择当前账号的配额，增加节点数量，否则会导致扩容失败。
 - 缩容时，您需要在操作列单击开关删除减少的节点。如图9-1所示，减少1个节点，需在“节点列表”中，单击删除节点对应操作列的开关，删除此节点。

图 9-1 缩容时选择删除节点

节点列表	名称	状态	创建时间	可用CPU	可用内存	可用GPU	操作
modelarts-155608694903...	运行中	2019/04/24 14:22:30 GMT+08...	7.6099997	28297.918 MB	0	<input checked="" type="checkbox"/>	
modelarts-155608694903...	运行中	2019/04/24 14:22:30 GMT+08...	7.6099997	28297.918 MB	0	<input checked="" type="checkbox"/>	

3. 单击“提交”完成修改。提交完成后系统自动返回专属资源池管理页面。

删除专属资源池

当AI业务开发不再需要使用专属资源池时，您可以删除专属资源池，释放资源，减少费用成本。



说明

- 专属资源池删除后，将导致使用此资源的训练作业、Notebook、部署上线等不可用，且删除后不可恢复，请谨慎操作。
- “包年/包月”类型的专属资源池，不支持删除操作。

1. 进入专属资源池管理页面，在专属资源池所在行，单击操作列“删除”。
2. 在弹出的确认对话框中，单击“确认”，完成资源删除。

10 权限管理

10.1 权限管理基本概念

如果您需要对ModelArts进行精细的权限管理，您可以使用统一身份认证服务（Identity and Access Management，简称IAM），通过IAM，您可以：

- 根据企业的业务组织，在您的华为云账号中，给企业中不同职能部门的员工创建IAM用户，让员工拥有唯一安全凭证，并使用ModelArts资源。
- 根据企业用户的职能，设置不同的访问权限，以达到用户之间的权限隔离。

如果华为云账号已经能满足您的要求，不需要创建独立的IAM用户，您可以跳过本章节，不影响您使用ModelArts服务的其它功能。

本手册写作使用IAM的常见操作，包括创建用户、用户组、给用户组授权以及创建自定义策略，如果需要使用IAM进行其它操作，请参见[《统一身份认证服务用户指南》](#)。

账号

当您首次使用华为云时，需要使用手机号注册一个账号，该账号是您的华为云资源归属、资源使用计费的主体，对其所拥有的资源及云服务具有完全的访问权限，可以重置用户密码、分配用户权限等。账号统一接收所有IAM用户进行资源操作时产生的费用账单。账号在登录华为云控制台时，使用“账号登录”方式登录。

如果您忘记了账号的登录密码，可以重置密码，重置方法请参见：[忘记账号密码](#)。

图 10-1 账号登录



IAM 用户

由账号在IAM中创建的用户，是云服务的使用人员，具有身份凭证（密码和访问密钥），可以使用自己单独的用户名和密码通过控制台或者API访问华为云。根据账号授予的权限，帮助账号管理资源。IAM用户不拥有资源，不进行独立的计费，这些用户的权限和资源由所属账户统一控制和付费。IAM用户在华为云控制台登录时，使用“IAM用户登录”方式登录。

如果您忘记了IAM用户的登录密码，可以重置密码，重置方法请参见：[IAM用户忘记密码](#)。

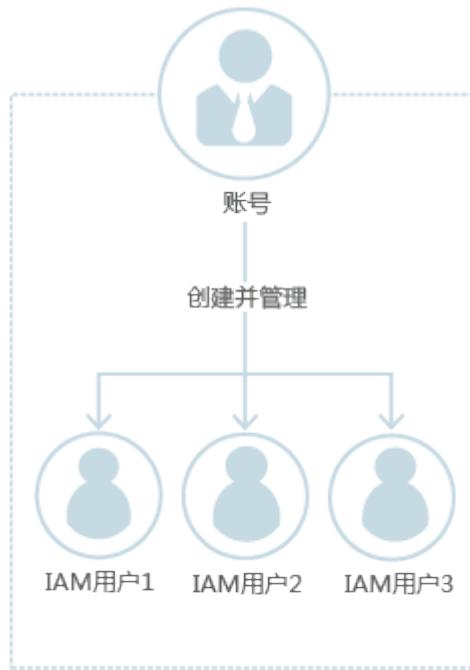
图 10-2 IAM 用户登录



账号与 IAM 用户的关系

账号与IAM用户可以类比为父子关系，账号是资源归属以及计费的主体，对其拥有的资源具有所有权限。IAM用户由账号创建，只能拥有账号授予的资源使用权限，账号可以随时修改或者撤销IAM用户的使用权限。IAM用户进行资源操作时产生的费用统一计入账号中，IAM用户不需要为资源付费。

图 10-3 账号与 IAM 用户的关系



身份凭证

身份凭证是识别用户身份的依据，您通过控制台或者API访问华为云时，需要使用身份凭证来通过系统的鉴权认证。身份凭证包括密码和访问密钥，您可以在IAM中管理自己以及账号中IAM用户的身份凭证。

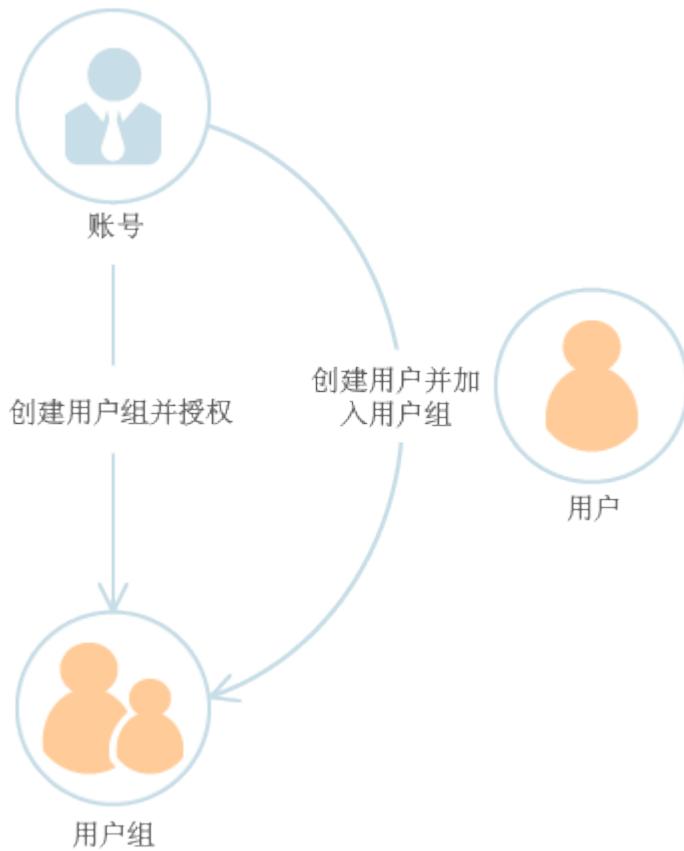
- 密码：常见的身份凭证，密码可以用来登录华为云界面控制台，还可以调用华为云的API接口。
- 访问密钥：即AK/SK（Access Key ID/Secret Access Key），调用华为云API接口的身份凭证，不能登录界面控制台。访问密钥中具有验证身份的签名，通过加密签名验证可以确保机密性、完整性和请求双方身份的正确性。

用户组

用户组是用户的集合，IAM通过用户组功能实现用户的授权。您在IAM中创建的用户，需要加入特定用户组后，用户才具备对应的权限，否则他们无法访问您账号中的任何资源或是云服务。一个用户可以加入多个用户组，以获得不同的权限。

“admin”为系统缺省提供的用户组，具有所有云服务资源的操作权限。将用户加入该用户组后，用户可以操作并使用所有云资源，包括但不限于创建用户组及用户、修改用户组权限、管理华为云云资源等。

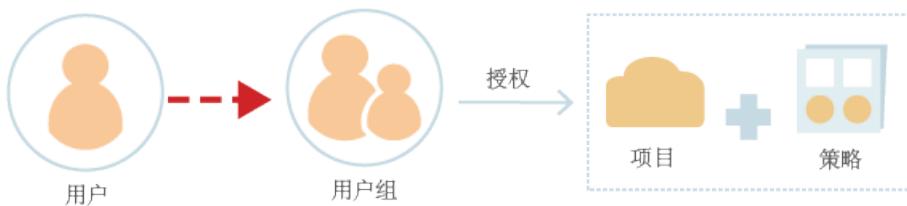
图 10-4 创建用户以及用户组



授权

授权是您将用户完成具体工作需要的权限授予用户，授权通过策略定义的权限生效，通过给用户组授予策略（包括系统策略和自定义策略），用户组中的用户就能获得策略中定义的权限，这一过程称为授权。用户获得具体云服务的权限后，可以对云服务进行操作，例如，管理您账号中的ECS资源。

图 10-5 授权模型

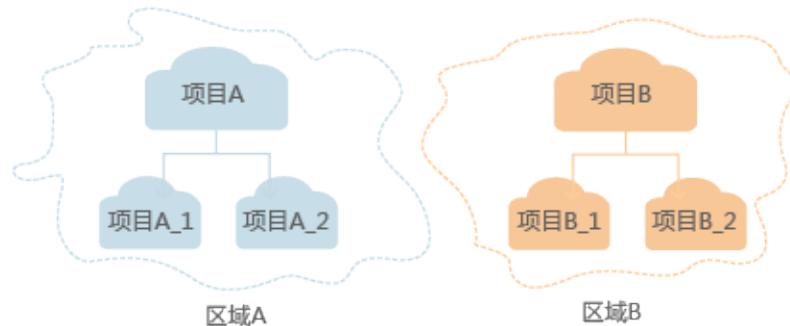


项目

华为云的区域默认对应一个项目，这个项目由系统预置，用来隔离物理区域间的资源（计算资源、存储资源和网络资源），以默认项目为单位进行授权，用户可以访问您

账号中该区域的所有资源。如果您希望进行更加精细的权限控制，可以在区域默认的项目中创建子项目，并在子项目中购买资源，然后以子项目为单位进行授权，使得用户仅能访问特定子项目中资源，使得资源的权限控制更加精确。

图 10-6 项目隔离模型



10.2 创建并授权使用 ModelArts

本章节通过简单的用户组授权方法，将ModelArts的策略授予用户组，并将用户添加至用户组中，从而使用户拥有对应的ModelArts权限，操作流程如图10-7所示。

示例流程

图 10-7 给用户授权 ModelArts 权限流程



1. 创建用户组并授权
在IAM控制台创建用户组，并授ModelArts User权限。
2. 创建用户

在IAM控制台创建用户，并将其加入**1**中创建的用户组。

3. 用户登录并验证权限

新创建的用户登录控制台，验证ModelArts的权限。

前提条件

- “ModelArts User权限”为细粒度策略，请先在IAM控制台中开通细粒度策略，开通方法请参见：[申请细粒度访问控制公测](#)。
- 给用户组授权之前，请您了解用户组可以添加的ModelArts系统策略，请参见[权限管理](#)。若您需要对除ModelArts之外的其它服务授权，IAM支持服务的所有策略请参见[权限策略](#)。

步骤1：创建用户组并授权

用户组是用户的集合，IAM通过用户组功能实现用户的授权。您在IAM中创建的用户，需要加入特定用户组后，用户才具备用户组所拥有的权限。关于创建用户组并给用户组授权的方法，可以参考如下操作。

步骤1 使用注册的华为云账号登录华为云，登录时请选择“账号登录”。



步骤2 进入华为云控制台，控制台页面中单击右上角的用户名，选择“统一身份认证”。



步骤3 在统一身份认证服务的左侧导航空格中，单击“用户组”>“创建用户组”。



步骤4 在“创建用户组”界面，输入“用户组名称”，以“开发人员组”为例，单击“确定”。

用户组创建完成，界面自动返回用户组列表，列表中显示新建的用户组。

步骤5 单击新建用户组右侧的“权限配置”，在“用户组权限”页签中，基于需要授权的区域，单击“设置策略”。

ModelArts为项目级服务，请确认用户需要使用ModelArts资源的项目，然后在对应项目中设置权限。

步骤6 在“设置策略”中搜索“ModelArts”，选择“ModelArts User”为例。ModelArts的系统策略说明，请参见[权限管理](#)。

步骤7 单击“确定”，完成用户组授权。

----结束

步骤 2：创建 IAM 用户

IAM用户与企业中的实际员工或是应用程序相对应，有唯一的安全凭证，可以通过加入一个或多个用户组来获得用户组的权限。关于IAM用户的创建方式请参见如下步骤。

步骤1 在统一身份认证服务，左侧导航中，单击“用户”>“创建用户”。

步骤2 在“创建用户”界面中填写参数信息，完成后单击“下一步”。

The screenshot shows the 'Create User' page. At the top, there are fields for '用户名' (Username) containing 'James' and '凭证类型' (Credential Type) with '密码' (Password) selected. Below this, there's a table for '所属用户组' (User Groups) showing '开发人员组' (Developer Group) assigned. A note says '适用于登录管理控制台，或者使用支持密码认证的API、CLI、SDK等开发工具来访问云服务。' (Applicable for logging into the management console or using password-authenticated APIs, CLIs, SDKs, etc. to access cloud services.) There's also a note for '访问密钥' (Access Key) which is not selected. Below the table is a text input for '请输入或选择用户组名称进行添加。' (Enter or select the user group name to add.) Under '描述' (Description), there's a text area with placeholder '请输入用户信息。' (Enter user information.) and a character limit of '0/255'. At the bottom are '下一步' (Next Step) and '取消' (Cancel) buttons.

- 用户名：用户登录华为云的用户名，以“James”为例。
- 凭证类型：凭证是指用户系统认证的身份凭证，以选择“密码”为例。
 - 密码：用户需要同时登录界面以及通过开发工具（API、CLI、SDK）访问华为云。
 - 访问密钥：用户仅需要通过开发工具访问华为云，不需要登录界面，凭证类型建议选择访问密钥，更加安全。
- 所属用户组（可选）：选择新创建的用户组“开发人员组”。将用户加入用户组，用户将具备用户组的权限，这一过程即给该用户授权。其中“admin”为系统缺省提供的用户组，具有管理人员以及所有云服务资源的操作权限。
- 描述（可选）：对用户的描述信息。

步骤3 在界面中填写参数信息，单击“确定”，完成用户创建。

密码生成方式：

- 首次登录时设置：如果您不是当前新建用户的使用主体，建议您选择该方式。用户通过邮件中的一次性链接登录华为云，自行设置密码。

- 自动生成：此用户是通过开发工具访问华为云，建议您选择该方式，华为云将自动生成随机的10位密码。
- 自定义：如果您是用户James的使用主体，建议您选择该方式，设置自己的登录密码。

----结束

步骤 3：用户登录并验证权限

用户创建完成后，可以使用新用户的用户名及身份凭证登录华为云验证权限，即“ModelArts User”权限。更多用户登录方法请参见[用户登录华为云方法](#)。

步骤1 在华为云登录页面，单击右下角的“IAM用户登录”。



步骤2 在“IAM用户登录”页面，输入账号名、用户名及用户密码，使用新创建的用户登录。

- 账号名为该IAM用户所属华为云账号的名称。
- 用户名和密码为账号在IAM创建用户时输入的用户名和密码。

如果登录失败，您可以联系您的账号主体，确认用户名及密码是否正确，或是重置用户名及密码，重置方法请参见：[忘记IAM用户密码](#)。

步骤3 登录成功后，进入华为云控制台，登录后默认区域为“华北-北京一”，请先切换至授权区域。



步骤4 进入ModelArts主界面，单击“专属资源>创建”，若提示权限不足，表示“ModelArts Viewer”已生效。

----结束

10.3 创建 ModelArts 自定义策略

如果系统预置的ModelArts权限，不满足您的授权要求，可以创建自定义策略。自定义策略中可以添加的授权项（Action）请参考[《ModelArts API参考》>权限策略和授权项](#)。如下以定制一个禁止创建和删除资源的权限策略为例。

前提条件

- 请先在IAM控制台中开通细粒度策略，开通方法请参见：[申请细粒度访问控制公测](#)。
- 自定义策略需要编写策略（JSON格式），请您先熟悉策略结构，具体请参见[策略语法：细粒度策略](#)。
- 请确定自定义策略需要允许哪些操作，拒绝哪些操作，并获取操作对应的授权项。授权项请参见：[《ModelArts API参考》>权限策略和授权项](#)。

操作步骤

如下以创建名为“modelarts_deny_delete_or_create”的策略为例，创建一个禁止创建和删除资源权限的自定义策略。

步骤1 在IAM控制台，单击左侧导航栏的“策略”，在右上角选择“创建自定义策略”。



步骤2 在“创建自定义策略”中，填写如下参数。

图 10-8 创建策略



- “策略名称”：填写“modelarts_deny_delete_or_create”。
- “作用范围”：根据服务的属性填写，ModelArts为项目级服务，选择“项目级服务”。
- “策略信息”：将如下内容拷贝至策略信息中，并单击“检验语法”。如下策略表示禁止创建和删除资源权限。

```
{  
    "Version": "1.1",  
    "Statement": [  
        {  
            "Action": [  
                "modelarts:*:*delete*",  
                "modelarts:*:*create*"  
            ],  
            "Effect": "Deny"  
        }  
    ]  
}
```

步骤3 单击“确定”，自定义策略创建成功。

步骤4 将新创建的自定义策略授予用户组，使得用户组中的用户不具备创建和删除资源权限。

步骤5 用户登录并验证自定义策略定义的权限：创建自动学习项目。

权限授予成功后，用户可以通过控制台以及REST API等多种方式验证。此处以登录控制台为例，介绍用户如何验证创建资源的权限。

1. 使用新创建的用户登录华为云，登录方法选择为“IAM用户登录”。
 - 账号名为该IAM用户所属华为云账号的名称。
 - 用户名和密码为账号在IAM创建用户时输入的用户名和密码。
2. 在ModelArts管理控制台，进行创建自动学习项目操作，系统显示“权限不足”，权限配置正确并已生效。

----结束

策略样例

- 示例1：拒绝用户删除自动学习项目

拒绝策略需要同时配合其他策略使用，否则没有实际作用。用户被授予的策略中，一个授权项的作用如果同时存在Allow和Deny，则遵循Deny优先。

如果您给用户授予ModelArts Admin的系统策略，但不希望用户拥有ModelArts Admin中定义的删除自动学习项目权限，您可以创建一条拒绝删除自动学习项目的自定义策略，然后同时将ModelArts Admin和拒绝策略授予用户，根据Deny优先原则，则用户可以对ModelArts执行除了删除自动学习项目外的所有操作。拒绝策略示例如下：

```
{  
    "Version": "1.1",  
    "Statement": [  
        {  
            "Effect": "Deny",  
            "Action": [  
                "modelarts:exmlProject:delete"  
            ]  
        }  
    ]  
}
```

- 示例2：仅为用户授予开发环境的使用权限

为某一用户配置ModelArts开发环境的使用权限时，由于依赖OBS服务的授权，因此必须配置相关的OBS最小化权限项，包含OBS桶和OBS对象的权限。建议采用[多个授权项策略](#)的方式配置，此用户的策略配置示例如下所示：

```
{  
    "Version": "1.1",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": [  
                "obs:bucket>ListAllMyBuckets",  
                "obs:bucket>CreateBucket",  
                "obs:bucket>ListBucket",  
                "obs:bucket>ListBucketVersions",  
                "obs:bucket>HeadBucket",  
                "obs:bucket>PutBucketAcl",  
                "obs:object>PutObject",  
                "obs:object>GetObject",  
                "obs:object>GetObjectVersion",  
                "obs:object>GetObjectVersionAcl"  
            ]  
        },  
        {  
            "Effect": "Allow",  
            "Action": [  
                "modelarts:notebook:list",  
                "modelarts:notebook:create",  
                "modelarts:notebook:get",  
                "modelarts:notebook:update",  
                "modelarts:notebook:delete",  
                "modelarts:notebook:action"  
            ]  
        }  
    ]  
}
```

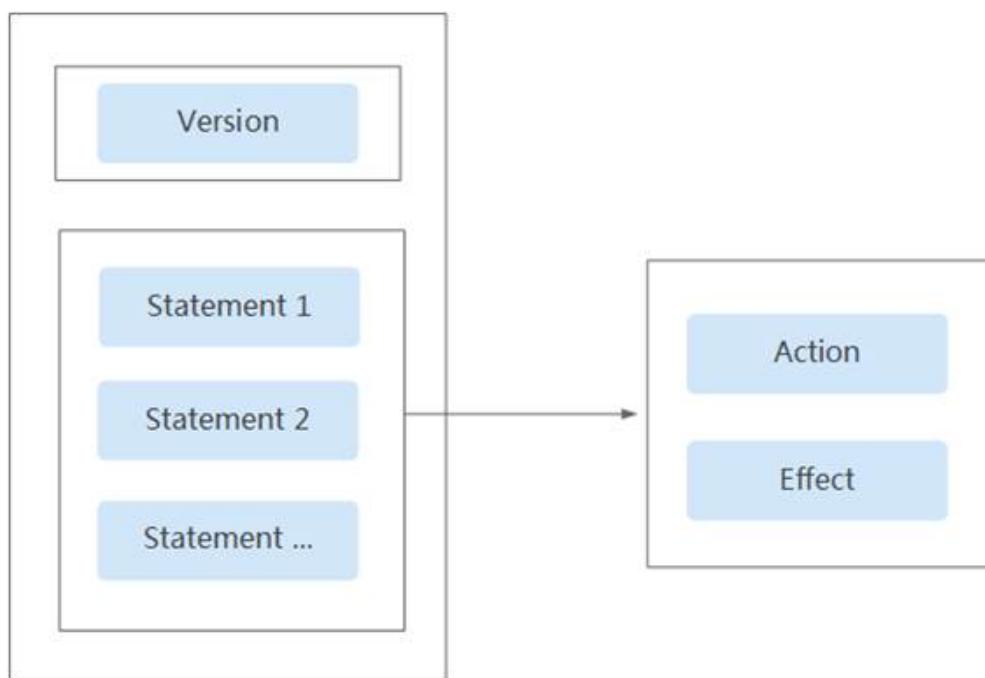
```
        "modelarts:notebook:access"
    ]
}
}
```

10.4 策略语法：细粒度策略

策略结构

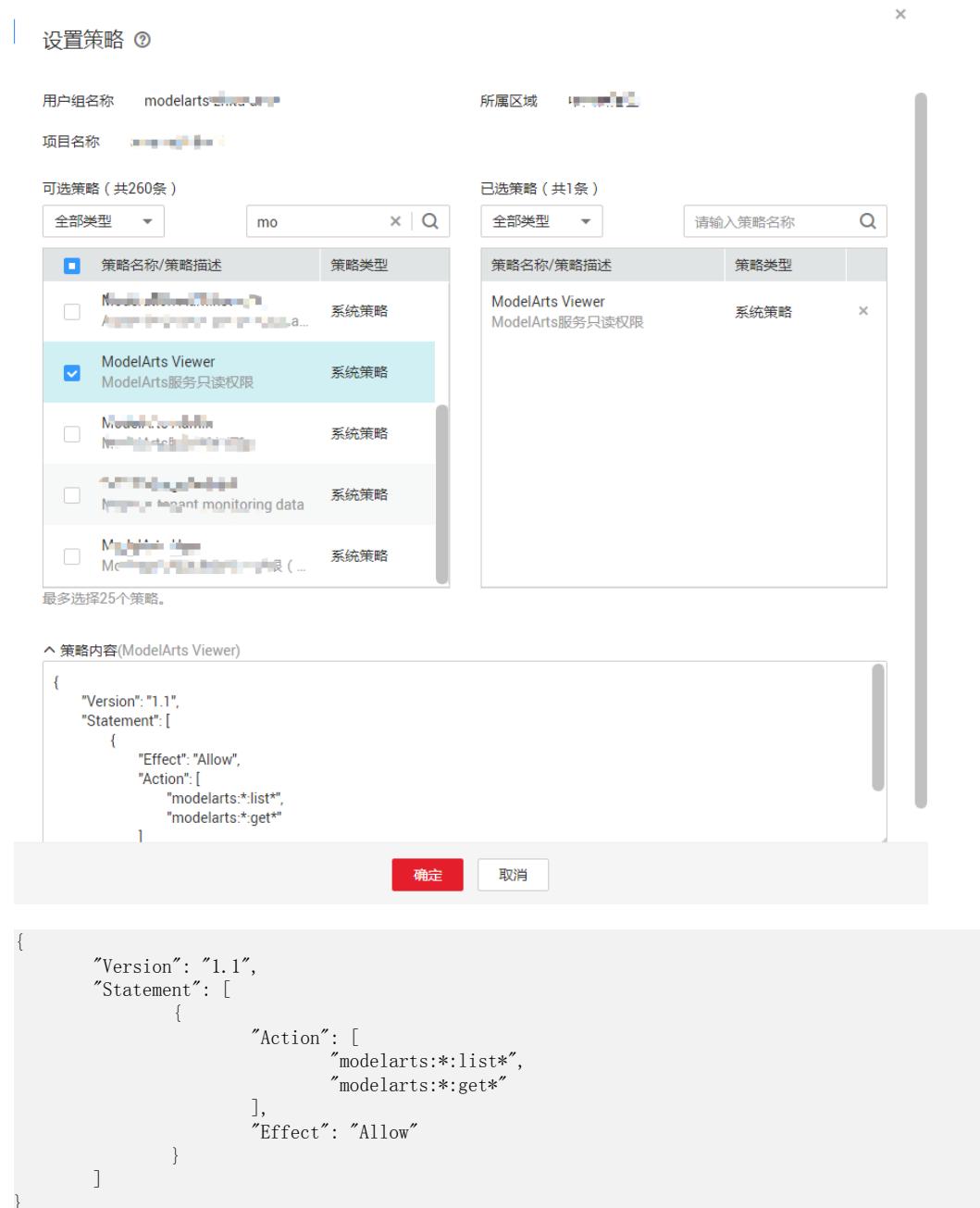
策略结构包括：Version（策略版本号）和Statement（策略权限语句），其中Statement可以有多个，表示不同的授权项。

图 10-9 策略结构



策略语法

如下以“ModelArts Viewer策略”为例，说明策略语法组成。



- **Version:** 标识策略的版本号，主要用于区分Role-Based Access Control (RBAC) 策略和细粒度策略。
 - 1.0: RBAC策略。RBAC策略是将服务作为一个整体进行授权，授权后，用户可以拥有这个服务的所有权限。
 - 1.1: 细粒度策略。相比RBAC策略，细粒度策略基于服务的API接口进行权限拆分，授权更加精细。授权后，用户可以对这个服务执行特定的操作。细粒度策略包括系统预置和用户自定义两种。
- **Statement:** 策略授权语句，描述策略的详细信息，包含Effect（作用）和Action（授权项）。
 - Effect（作用）

作用包含两种：Allow（允许）和Deny（Deny），系统预置策略仅包含允许的授权语句，自定义策略中可以同时包含允许和拒绝的授权语句，当策略中既有允许又有拒绝的授权语句时，遵循Deny优先的原则。

- Action（授权项）

对资源的具体操作权限，格式为：服务名:资源类型:操作，支持单个或多个操作权限，支持通配符号*，通配符号表示所有。

示例："modelarts:exemlProject:create"，其中modelarts为服务名，exemlProject为项目类型，create为操作，该授权项表示ModelArts创建自动学习项目权限。

多个授权项策略

一个自定义策略中可以包含多个授权项，且除了可以包含本服务的授权项外，还可以包含其他服务的授权项，可以包含的其他服务必须跟本服务同属性，即都是项目级服务。多个授权语句策略描述如下：

```
{  
    "Version": "1.1",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": [  
                "ecs:cloudServers:resize",  
                "ecs:cloudServers:delete",  
                "ecs:cloudServers:rebuild"  
            ]  
        },  
        {  
            "Effect": "Allow",  
            "Action": [  
                "modelarts:exemlProjectVersion:delete",  
                "modelarts:exemlProjectVersion:delete"  
            ]  
        }  
    ]  
}
```

拒绝策略

拒绝策略需要同时配合其他策略使用，否则没有实际作用。用户被授予的策略中，一个授权项的作用如果同时存在Allow和Deny，则遵循Deny优先。

如果您给用户授予ModelArts Admin的系统策略，但不希望用户拥有ModelArts Admin中定义的删除自动学习项目权限，您可以创建一条拒绝删除自动学习项目的自定义策略，然后同时将ModelArts Admin和拒绝策略授予用户，根据Deny优先原则，则用户可以对ModelArts执行除了删除自动学习项目外的所有操作。拒绝策略示例如下：

```
{  
    "Version": "1.1",  
    "Statement": [  
        {  
            "Effect": "Deny",  
            "Action": [  
                "modelarts:exemlProject:delete"  
            ]  
        }  
    ]  
}
```

检查规则

当用户被授予多个策略，或者一个策略中包含多个授权语句时，遵循Deny优先的原则。在用户访问资源时，权限检查逻辑如下。

图 10-10 系统鉴权逻辑图



说明

每条策略做评估时，Action之间是或(or)的关系。

1. 用户访问系统，发起操作请求。
2. 系统评估用户被授予的访问策略，鉴权开始。
3. 在用户被授予的访问策略中，系统将优先寻找显式拒绝指令。如找到一个适用的显式拒绝，系统将返回Deny决定。
4. 如果没有找到显式拒绝指令，系统将寻找适用于请求的任何Allow指令。如果找到一个显式允许指令，系统将返回Allow决定。
5. 如果找不到显式允许，最终决定为Deny，鉴权结束。

10.5 策略语法：RBAC

策略结构

策略结构包括：策略版本号（Version）、策略授权语句（Statement）和策略依赖（Depends）。

图 10-11 策略结构



策略语法

如下以SDRS服务的“SDRSAdministrator”为例，说明RBAC策略语法。



```
{  
    "Version": "1.0",  
    "Statement": [  
        {  
            "Action": [  
                "SDRS:*:*"  
            ],  
            "Effect": "Allow"  
        }  
    ],  
    "Depends": [  
        {  
            "catalog": "BASE",  
            "display_name": "Tenant Guest"  
        },  
        {  
            "catalog": "BASE",  
            "display_name": "Server Administrator"  
        }  
    ]  
}
```

参数	含义	值
Version	策略的版本	固定为“1.0”

参数		含义	值
Statement	Action	定义对SDRS的具体操作。	格式为：服务名:资源类型:操作 "SDRS:*:*", 表示对SDRS的所有操作，其中SDRS为服务名称；“*”为通配符，表示对所有的资源类型可以执行所有操作。
	Effect	定义Action中所包含的具体操作是否允许执行。	<ul style="list-style-type: none">● Allow: 允许执行。● Deny: 不允许执行。
Depends	catalog	依赖的其他策略的所属目录。	服务名称 例如：BASE
	display_name	依赖的其他权限的名称。	权限名称 例如：Tenant Administrator

11 使用自定义镜像

11.1 自定义镜像简介

ModelArts提供了多种预置引擎，但是当用户对深度学习引擎、开发库有特殊需求的场景的时候，预置AI引擎已经不能满足用户需求。此时用户可以使用ModelArts自定义镜像这个功能来使用自定义运行引擎。

ModelArts底层采用容器技术，自定义镜像指的是用户自行制作容器镜像并在ModelArts上运行。自定义镜像功能支持自由文本形式的命令行参数和环境变量，因此灵活性比较高，便于支持任意计算引擎的作业启动需求。

制作自定义镜像还需要使用的华为云服务有：容器镜像服务SWR、对象存储服务OBS、弹性云服务器ECS。

自定义镜像流程

ModelArts中使用自定义功能步骤如下：

1. 制作自定义镜像。
2. 上传自定义镜像到SWR。
3. 在ModelArts中使用自定义镜像。

自定义镜像规范

- 镜像对外端口

镜像的对外服务端口需要为8080，访问PATH需要为“/”，当镜像启动时可以直接访问。下面是mnist镜像的访问示例，该镜像内含mnist数据集训练的模型，可以识别手写数字。

- 请求示例`curl -X POST \ http://{listen_ip}:8080/ \ -F images=@seven.jpg`
- 返回示例
`{"mnist_result": 7}`

- 健康检查端口

自定义镜像需要提供健康检查接口供ModelArts调用，健康检查接口规范如下。

- URI
`GET /health`

- 请求示例`curl -X GET \http://{listen_ip}:8080/health`
- 响应示例

```
{"health": "true"}
```
- 状态码

表 11-1 状态码

状态码	编码	状态码说明
200	OK	请求成功

- 日志文件输出

为保证日志内容可以正常显示，日志信息需要打印到标准输出。

- 镜像启动入口

如果需要部署批量服务，镜像的启动入口文件需要为“/home/run.sh”，采用CMD设置默认启动路径，例如Dockerfile如下：

CMD /bin/sh /home/run.sh

- 镜像依赖组件

如果需要部署批量服务，镜像内需要安装python、jre/jdk、zip等组件包。

11.2 制作自定义镜像

制作自定义镜像时，需要往镜像中添加一些必要的深度学习库及用户编写的脚本等。

约束限制

- 自定义镜像必须基于ModelArts官方提供的基础镜像。
- 自定义镜像中不能包含恶意代码。
- 基础镜像中的部分内容不能改变，包括“/bin”、“/sbin”、“/usr”、“/lib(64)”下的所有文件，“/etc”下的部分重要配置文件，以及“\$HOME”下的ModelArts小工具。
- 不得新增属主为“root”且权限包含“setuid”或“setgid”位的文件。
- 自定义镜像不能超过9.5GB。

基础镜像介绍

基础镜像中有一些必要的工具，用户需要基于ModelArts官方提供的基础镜像来制作自定义镜像。

基础镜像基础组件如下所示。

- `run_train.sh`: 训练启动脚本，调用加解密工具，下载代码目录，日志输出，执行训练命令。
- `dls-key-client`、`dls-decryptor`: 加解密工具。
- `dls-dns-fixer`: DNS工具。
- `dls-pipe`: 日志工具。
- `dls-downloader`: OBS下载工具。

- ip-mapper: IP转换工具。

各个基础镜像包含的其他公共库，请参见[Dockerfile](#)。

ModelArts会不断更新基础镜像，基础镜像更新后，对于兼容性更新，用户还可以继续使用旧的镜像；对于不兼容性更新，基于旧版本制作的自定义镜像将不能在ModelArts上运行，但已经审核过的自定义镜像可以继续使用。

当用户发现自定义镜像审核不通过，并且审核日志中报镜像不匹配的时候，建议更新基础镜像。附录中会列出各个基础镜像。

制作自定义镜像

您可以通过如下两种方式制定自定义镜像：

1. 编写[Dockerfile](#)文件，采用SWR自动构建镜像功能，详细指导请参见[构建镜像](#)。
2. 使用自己的电脑搭建Docker环境，或者在华为云上的[购买1台ECS](#)来搭建Docker环境。参考[SWR使用指导](#)将基础镜像pull到本地，并制作自定义镜像。

推荐使用第一种方法，这种方法比较简单，但是需要有Dockerfile基础。

11.3 上传镜像至 SWR

用户制作完成镜像后，请参考[上传镜像至容器镜像服务](#)将镜像上传到自己的SWR中。

11.4 在 ModelArts 中使用自定义镜像

创建自定义镜像训练用户作业

在ModelArts管理控制台，使用自定义镜像创建训练作业。

● 算法来源说明

如图11-1所示，在“算法来源”中，选择“自定义”。

- “镜像地址”：镜像上传到SWR后生成的地址。
- “代码目录”：用户存放训练代码的OBS路径（非必须）。
- “运行命令”：镜像启动后的运行命令，基本格式如下所示。

`bash /home/work/run_train.sh {UserCommand}`

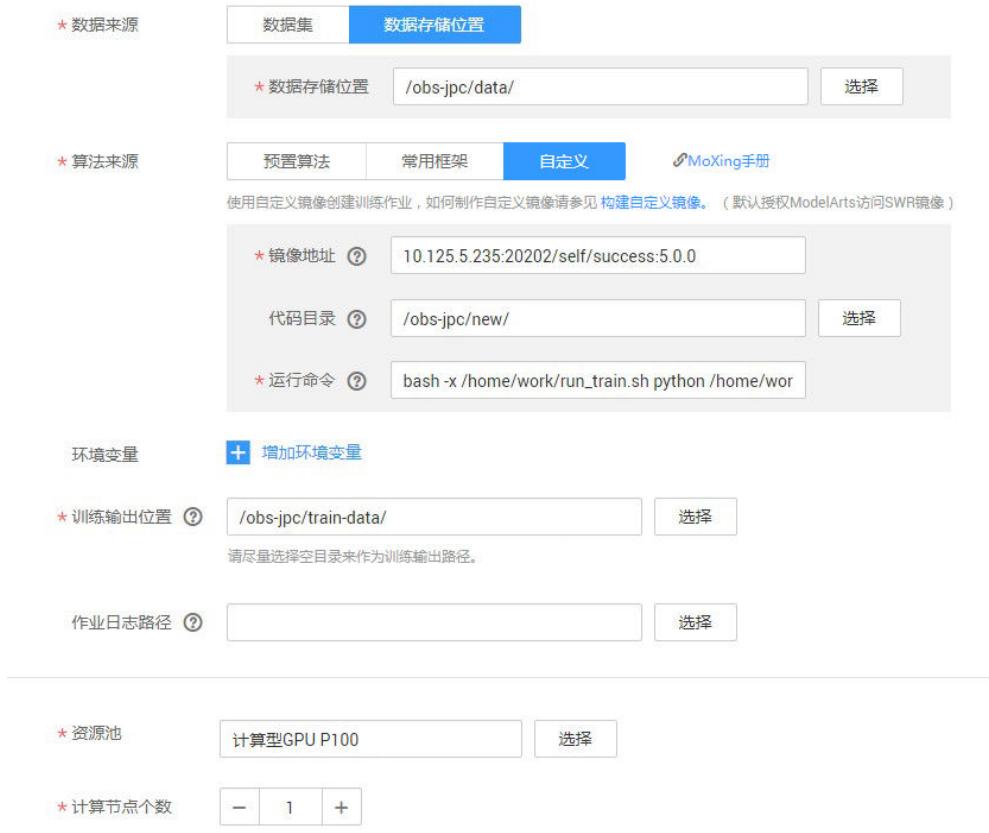
`bash /home/work/run_train.sh [python/bash/..] {file_location} {file_parameter}`

采用这种形式，“代码目录”配置后，“run_train.sh”将代码目录下载到容器的“/home/work/user-job-dir”目录中。

例如，训练代码启动脚本在OBS路径为“s3://obs/app/train.py”，创建作业时选择OBS代码路径为“/obs/app/”，app目录会下载到“/user-job-dir”下，运行命令就可以设置为：

`bash /home/work/run_train.sh python /home/work/user-job-dir/app/train.py {python_file_parameter}`

图 11-1 创建训练作业



● 环境变量说明

容器启动后，除了用户在训练作业中自行增加的“环境变量”外，其它加载的环境变量如**表11-2**所示。用户可以根据需求来确认在自己训练脚本的python中要不要使用这些环境变量，也可以通过运行命令中的“{python_file_parameter}”传入相关参数。

表 11-2 可选环境变量说明

环境变量	说明
DLS_TASK_INDEX	当前容器索引，容器从0开始编号。单机训练的时候，该字段无意义。在分布式作业中，用户可以根据这个值来确定当前容器运行的算法逻辑。
DLS_TASK_NUMBER	容器总数。对应“计算节点个数”。
DLS_APP_URL	代码目录。对应界面上“代码目录”配置，会加上协议名。比如，可直接使用“\$DLS_APP_URL/*.py”来读取OBS下的文件。
DLS_DATA_URL	数据集位置。对应界面上“数据来源”，会加上协议名。
DLS_TRAIN_URL	训练输出位置。对应界面上“训练输出位置”，会加上协议名。

环境变量	说明
BATCH_CUSTOM"\${i}"_HOSTS (分布式)	当选择分布式的时候才有这个环境变量，即计算节点个数大于1时，此环境变量为“BATCH_CUSTOM"\${i}"_HOSTS”。
BATCH_{jobName}.0_HOSTS (单机)	当选择单机时，即计算节点个数为1时，此环境变量为“BATCH_{jobName}.0_HOSTS”。 容器网络，容器的“hostname:port”。一个容器可以看到同一个作业中所有容器的HOSTS，根据索引的不同，分别为“BATCH_CUSTOM0_HOSTS”、“BATCH_CUSTOM1_HOSTS”等。当前，当资源池为8卡专属池的时候，容器为主机网络启动，并且可以使用主机IB网络加速通信；其他资源池为容器网络。

● 资源池说明

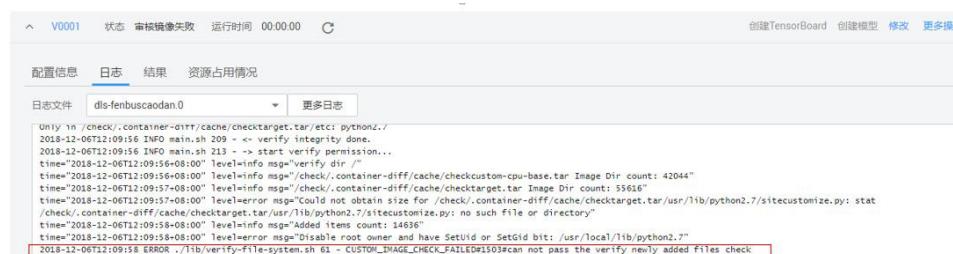
当用户选择GPU时，默认会挂载缓存到“/cache”目录，用户可以使用此目录来储存临时文件。

自定义镜像让用户对使用ModelArts有最大的灵活性，用户可以不采用**bash /home/work/run_train.sh {command}**形式，把所需的东西全放在镜像，直接让容器执行某些指令。采用**bash /home/work/run_train.sh {command}**形式，用户可以随意调参。

运行自定义镜像训练作业

用户上传自定义镜像到SWR后，在创建自定义镜像作业时，默认已经授权 ModelArts去获取镜像运行。自定义审核镜像第一次运行的时候，先审核镜像，审核内容请参见[自定义镜像规范](#)，审核失败的原因见于日志，用户根据日志做相应的修改。

图 11-2 审核镜像失败



镜像审核成功后，后台就会开始启动用户自定义镜像容器，开始跑自定义镜像训练作业，用户可根据日志来查看训练情况。

图 11-3 运行日志



审核成功后，再次使用相同镜像创建训练作业的时候，不会再次审核。

11.5 自定义镜像创建训练作业配置示例

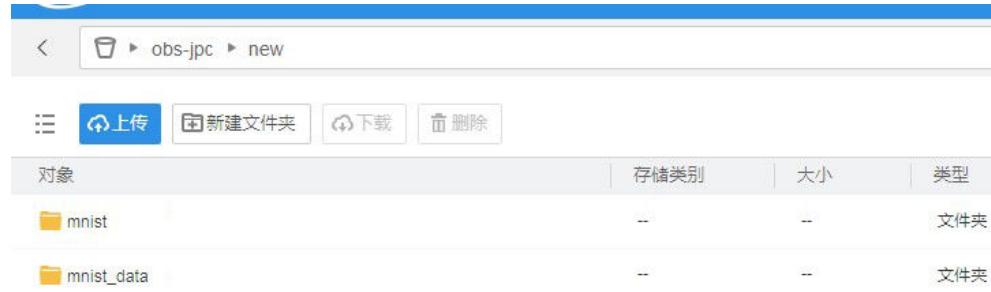
本示例所需的文件存储在[Github 目录](#)中。例子中的分布式和单机的区别主要在于脚本和运行命令不一样，使用的镜像都为同一个。

本示例使用MNIST训练数据，数据集可从[MNIST官网](#)下载。“mnist_softmax.py”为单机脚本，“mnist_replica_kill.py”为分布式脚本。

单机训练示例

1. 下载基础镜像，安装Tensorflow深度学习库，然后把镜像push到SWR。也可以采用[Dockerfile](#)在SWR上构建。
2. 将“mnist_softmax.py”和训练数据上传至OBS。现将脚本和数据都放在代码目录下，以便直接下载到容器中。

图 11-4 上传训练脚本和数据



3. 创建自定义镜像训练作业，数据存储位置和训练输出位置请根据实际情况填写。

- “镜像地址”：填写刚上传镜像的“SWR_URL”。
- “运行命令”：

bash /home/work/run_train.sh python /home/work/user-job-dir/new/mnist/mnist_softmax.py --data_url /home/work/user-job-dir/new/mnist_data

其中，“/home/work/user-job-dir/new/mnist/mnist_softmax.py”为下载下来训练脚本的位置，“--data_url /home/work/user-job-dir/new/mnist_data”为数据的位置。由于已经把数据放在代码目录中，容器已经下载了代码目录，所以直接使用本地的。

4. 自定义镜像审核成功后，后台会直接执行自定义镜像训练作业。程序执行成功后，日志信息如下所示。

图 11-5 运行日志

```
tensorflow.contrib.learn.python.learn.datasets.mnist is deprecated and will be removed in a future version.  
Instructions for updating:  
Please use alternatives such as official/mnist/dataset.py from tensorflow/models.  
job name = ps  
task index = 0  
2018-12-13 14:19:56.438111: I tensorflow/core/platform/cpu_feature_guard.cc:140] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX2 FMA  
2018-12-13 14:19:56.442049: I tensorflow/core/distributed_runtime/rpc/grpc_channel.cc:215] Initialize GrpcChannelCache for job ps -> {0 -> localhost:6666}  
2018-12-13 14:19:56.442077: I tensorflow/core/distributed_runtime/rpc/grpc_channel.cc:215] Initialize GrpcChannelCache for job worker -> {0 -> jobhnzmnro-custom1-0:6667}  
2018-12-13 14:19:56.447842: I tensorflow/core/distributed_runtime/rpc/grpc_server_lib.cc:332] Started server with target: grpc://localhost:6666  
2018-12-13 14:19:57.467433: I tensorflow/core/distributed_runtime/master_session.cc:1136] Start master session 77b9942900e3e38a6 with config:  
ps 0 received done 0
```

分布式训练示例

1. 分布式例子和单机的例子不同的点在于需要修改“python”文件。对“DLS_TASK_INDEX”和“DLS_TASK_NUMBER”进行处理，来适配脚本所需参数，决定当前容器的功能。当前脚本的功能是将前两个容器作为“ps”，后两个容器作为“worker”。

图 11-6 脚本修改



```

48 task_index = os.getenv('DLS_TASK_INDEX')
49 task_num = os.getenv('DLS_TASK_NUMBER')
50
51 #if task number big then 1, it is a distribution task
52 #half is ps, else is worker
53 if int(task_num) > 1:
54     batch_ps_hosts = ''
55     batch_worker_hosts = ''
56     middle_index = int(task_num) / 2
57     for i in range(0, int(task_num)):
58         print ('{:d}'.format(i))
59         if i < middle_index:
60             batch_ps_hosts += os.getenv('BATCH_CUSTOM' + str(i) + '_HOSTS') + ','
61         else:
62             batch_worker_hosts += os.getenv('BATCH_CUSTOM' + str(i) + '_HOSTS') + ','
63
64     batch_ps_hosts= batch_ps_hosts[:-1]
65     batch_worker_hosts= batch_worker_hosts[:-1]
66     job_name = ''
67     if int(task_index) < middle_index:
68         job_name = 'ps'
69     else:
70         job_name = 'worker'
71     task_index = str(int(int(task_index) - middle_index))
72
73     print('job name:' + job_name)
74     print('job index:' + task_index)
75     print('batchPsHosts:' + batch_ps_hosts)
76     print('batchWorkerHosts:' + batch_worker_hosts)
77     tf_task_index = int(task_index)
78
79
80     flags = tf.app.flags
81     flags.DEFINE_string("data_url", "/tmp/mnist-data",
82                         "Directory for storing mnist data")
83     flags.DEFINE_boolean("download_only", False,
84                          "Only perform downloading of data; Do not proceed to "
85                          "session preparation, model definition or training")
86     flags.DEFINE_integer("task_index", tf_task_index,
87                          "Worker task index, should be >= 0. task_index=0 is "
88                          "the master worker task the performs the variable "
89                          "initialization")
90     flags.DEFINE_integer("num_gpus", 1,
91                          "Total number of gpus for each machine."
92                          "If you don't use GPU, please set it to '0'")
93
94
95

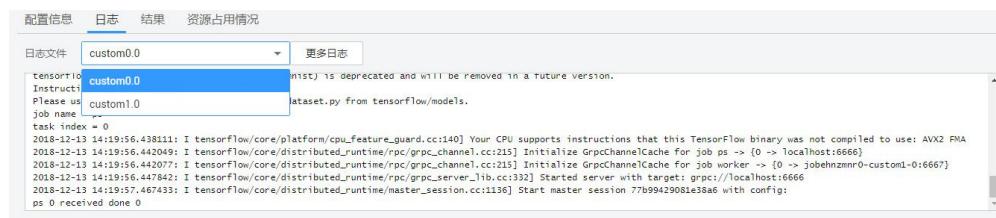
```

其中，运行命令为：

bash /home/work/run_train.sh python /home/work/user-job-dir/new/mnist/mnist_replica_kill.py --data_url /home/work/user-job-dir/new/mnist_data

2. 计算节点个数选择多个（即至少大于1），其他设置与单机一样。执行结果如下所示。

图 11-7 运行结果日志



配置信息	日志	结果	资源占用情况
日志文件	custom0	更多日志	
<pre> tensorFlow Instruct: custom0 [DEPRECATION] is deprecated and will be removed in a future version. Please use custom1_0 dataset.py from tensorflow/models. job name: '' task index: 0 2018-12-13 14:19:56.438111: I tensorflow/core/platform/cpu_feature_guard.cc:1140] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX2 FMA 2018-12-13 14:19:56.442049: I tensorflow/core/distributed_runtime/rpc/grpc_channel.cc:215] Initialize GrpcChannelCache For job ps -> [0 -> jobhnzmn0-custom1-0:6667] 2018-12-13 14:19:56.442077: I tensorflow/core/distributed_runtime/rpc/grpc_channel.cc:215] Initialize GrpcChannelCache For job worker -> [0 -> jobhnzmn0-custom1-0:6667] 2018-12-13 14:19:56.447842: I tensorflow/core/distributed_runtime/rpc/grpc_server_lib.cc:332] Started server with target: grpc://localhost:6669 2018-12-13 14:19:57.467433: I tensorflow/core/distributed_runtime/master_session.cc:1136] Start master session 77b99429081e38a6 with config: ps 0 received done </pre>			

11.6 附录：自定义镜像的补充说明

基础镜像列表

根据个人需求选择对应的GPU、CUDA、MoXing和Python版本。

```
swr.cn-north-1.myhuaweicloud.com/eiwizard/custom-cpu-base:1.2
swr.cn-north-1.myhuaweicloud.com/eiwizard/custom-gpu-cuda92-base:1.2
swr.cn-north-1.myhuaweicloud.com/eiwizard/custom-gpu-cuda9-base:1.2
swr.cn-north-1.myhuaweicloud.com/eiwizard/custom-gpu-cuda8-base:1.2
swr.cn-north-1.myhuaweicloud.com/eiwizard/custom-cpu-inner-moxing-cp27:1.2
swr.cn-north-1.myhuaweicloud.com/eiwizard/custom-gpu-cuda8-inner-moxing-cp27:1.2
swr.cn-north-1.myhuaweicloud.com/eiwizard/custom-gpu-cuda9-inner-moxing-cp27:1.2
swr.cn-north-1.myhuaweicloud.com/eiwizard/custom-gpu-cuda9-inner-moxing-cp36:1.2
```

模型训练的 SDK 或 API

如果您不想使用console的功能创建训练作业，您还可以调用SDK或者API实现。SDK的详细指导请参见《[ModelArts SDK参考](#)》。API的详细指导请参见《[ModelArts API参考](#)》。

12 模型包规范

12.1 模型包规范介绍

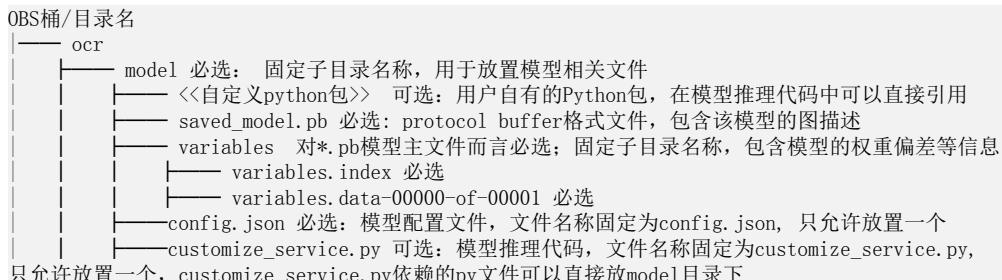
在模型管理导入模型时，如果是从OBS或容器镜像中导入元模型，则需要符合一定的模型包规范。

- 模型包里面必须包含“model”文件夹，“model”文件夹下面放置模型文件，模型配置文件，模型推理代码。
- 模型配置文件必需存在，文件名固定为“config.json”，有且只有一个，模型配置文件编写请参见[模型配置文件编写说明](#)。
- 模型推理代码文件可选，如果需要文件名固定为“customize_service.py”，且只能有一个，模型推理代码编写请参见[模型推理代码编写说明](#)。

模型包示例

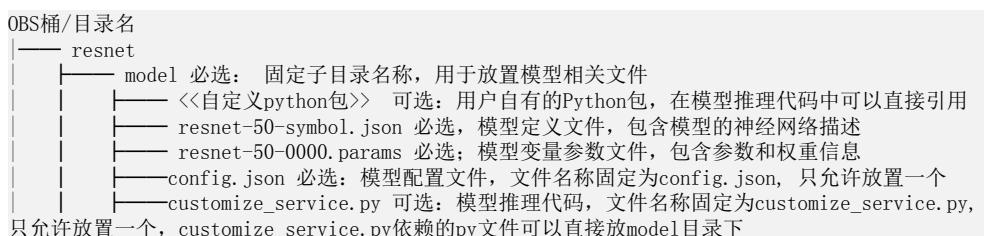
- TensorFlow模型包结构

发布该模型时只需要指定到“ocr”目录。



- MXNet模型包结构

发布该模型时只需要指定到“resnet”目录。



- **Image模型包结构**

发布该模型时只需要指定到“resnet”目录。

OBS桶/目录名

```
└── resnet
    ├── model 必选: 固定子目录名称, 用于放置模型相关文件
    └── config.json 必选: 模型配置文件(需要配置swr镜像地址), 文件名称固定为config.json, 只允许放置一个
```

- **pyspark模型包结构**

发布该模型时只需要指定到“resnet”目录。

OBS桶/目录名

```
└── resnet
    ├── model 必选: 固定子目录名称, 用于放置模型相关文件
    │   ├── <<自定义Python包>> 可选: 用户自有的Python包, 在模型推理代码中可以直接引用
    │   ├── spark_model 必选: 模型文件夹, 包含pyspark保存的模型内容
    │   ├── config.json 必选: 模型配置文件, 文件名称固定为config.json, 只允许放置一个
    │   └── customize_service.py 可选: 模型推理代码, 文件名称固定为customize_service.py, 只允许放置一个, customize_service.py依赖的py文件可以直接放model目录下
```

- **PyTorch模型包结构**

发布该模型时只需要指定到“resnet”目录。

OBS桶/目录名

```
└── resnet
    ├── model 必选: 固定子目录名称, 用于放置模型相关文件
    │   ├── <<自定义Python包>> 可选: 用户自有的Python包, 在模型推理代码中可以直接引用
    │   ├── resnet50.pth 必选, pytorch模型保存文件, 保存为“state_dict”, 存有权重变量等信息
    │   ├── config.json 必选: 模型配置文件, 文件名称固定为config.json, 只允许放置一个
    │   └── customize_service.py 可选: 模型推理代码, 文件名称固定为customize_service.py, 只允许放置一个, customize_service.py依赖的py文件可以直接放model目录下
```

- **Caffe模型包结构**

用户发布该模型时只需要指定到“resnet”目录

OBS桶/目录名

```
└── resnet
    ├── model 必选: 固定子目录名称, 用于放置模型相关文件
    │   ├── <<自定义python包>> 可选: 用户自有的Python包, 在模型推理代码中可以直接引用
    │   ├── deploy.prototxt 必选, caffe模型保存文件, 存有模型网络结构等信息
    │   ├── resnet.caffemodel 必选, caffe模型保存文件, 存有权重变量等信息
    │   ├── config.json 必选: 模型配置文件, 文件名称固定为config.json, 只允许放置一个
    │   └── customize_service.py 可选: 模型推理代码, 文件名称固定为customize_service.py, 只允许放置一个, customize_service.py依赖的py文件可以直接放model目录下
```

12.2 模型配置文件编写说明

模型开发者发布模型时需要编写配置文件。模型配置文件描述模型用途、模型计算框架、模型精度、推理代码依赖包以及模型对外API接口。

配置文件格式说明

配置文件为JSON格式，参数说明如[表12-1](#)所示。

表 12-1 参数说明

参数	是否必选	参数类型	描述
model_algorithm	是	String	模型算法，表明该模型的用途，由模型开发者填写，以便使用者理解该模型的用途，例如image_classification、object_detection。
model_type	是	String	模型AI引擎，表明模型使用的计算框架，可选的框架有TensorFlow、MXNet、Spark_MLlib、AutoModel、Scikit_Learn、XGBoost、Image、PyTorch。 Image镜像制作规范可参见 自定义镜像制作规范 。
runtime	否	String	模型运行时环境，可选“python2.7”、“python3.6”。
swr_location	否	String	SWR镜像模板地址。当“model_type”为“Image”，“swr_location”必填，“swr_location”为docker镜像在swr上的模板地址，表示直接使用SWR的docker镜像发布模型。
metrics	否	object数据结构	模型的精度信息，包括平均数、召回率、精确率、准确率，metrics object数据结构说明如 表12-2 所示。
apis	是	api数据结构数组	表示模型可对外提供的restful api数组，api数据结构如 表12-3 所示。
dependencies	否	dependency结构数组	推理代码及模型依赖的包，模型开发者需要提供包名、安装方式、版本约束，dependency结构数组说明如 表12-6 所示。

表 12-2 metrics object 数据结构说明

参数	是否必选	参数类型	描述
f1	否	Number	平均数。精确到小数点后17位，超过17位时，取前17位数值。
recall	否	Number	召回率。精确到小数点后17位，超过17位时，取前17位数值。
precision	否	Number	精确率。精确到小数点后17位，超过17位时，取前17位数值。
accuracy	否	Number	准确率。精确到小数点后17位，超过17位时，取前17位数值。

表 12-3 api 数据结构说明

参数	是否必选	参数类型	描述
protocol	是	String	请求协议。
url	是	String	请求路径。
method	是	String	请求方法。
request	是	Object	请求体, request结构说明如 表12-4 所示。
response	是	Object	响应体, response结构说明如 表12-5 所示。

表 12-4 request 结构说明

参数	是否必选	参数类型	描述
Content-type	是	String	data以指定内容类型发送。
data	是	String	请求体以json schema描述。

表 12-5 response 结构说明

参数	是否必选	参数类型	描述
Content-type	是	String	data以指定内容类型发送。
data	是	String	响应体以json schema描述。

表 12-6 dependency 结构数组说明

参数	是否必选	参数类型	描述
installer	是	String	安装方式, 当前只支持“pip”。
packages	是	package结构数组	依赖包集合, package结构数组说明如 表12-7 所示。

表 12-7 package 结构数组说明

参数	是否必选	参数类型	描述
package_name	是	String	依赖包名称。

参数	是否必选	参数类型	描述
package_version	否	String	依赖包版本，如果不强依赖于版本号，则该项不填。
restraint	否	String	版本限制条件，当且仅当“package_version”存在时必填，可选“EXACT/ATLEAST/ATMOST”。 <ul style="list-style-type: none">● “EXACT” 表示安装给定版本。● “ATLEAST” 表示安装版本不小于给定版本。● “ATMOST” 表示安装包版本不大于给定版本。

配置文件示例

- 目标检测模型配置文件示例

```
{  
    "model_type": "TensorFlow",  
    "model_algorithm": "object_detection",  
    "metrics": {  
        "f1": 0.345294,  
        "accuracy": 0.462963,  
        "precision": 0.338977,  
        "recall": 0.351852  
    },  
    "apis": [ {  
        "protocol": "http",  
        "url": "/",  
        "method": "post",  
        "request": {  
            "Content-type": "multipart/form-data",  
            "data": {  
                "type": "object",  
                "properties": {  
                    "images": {  
                        "type": "file"  
                    }  
                }  
            }  
        }  
    },  
    "response": {  
        "Content-type": "multipart/form-data",  
        "data": {  
            "type": "object",  
            "required": [  
                "detection_classes",  
                "detection_boxes",  
                "detection_scores"  
            ],  
            "properties": {  
                "detection_classes": {  
                    "type": "array",  
                    "item": [{  
                        "type": "string"  
                    }]  
                },  
                "detection_boxes": {  
                    "type": "array",  
                }  
            }  
        }  
    }  
}
```

```
        "items": [
            {
                "type": "array",
                "minItems": 4,
                "maxItems": 4,
                "items": [
                    {
                        "type": "number"
                    }
                ]
            }
        ],
        "detection_scores": {
            "type": "number"
        }
    }
}
],
"dependencies": [
    {
        "installer": "pip",
        "packages": [
            {
                "restraint": "ATLEAST",
                "package_version": "1.15.0",
                "package_name": "numpy"
            },
            {
                "restraint": "",
                "package_version": "",
                "package_name": "h5py"
            },
            {
                "restraint": "ATLEAST",
                "package_version": "1.8.0",
                "package_name": "tensorflow"
            },
            {
                "restraint": "ATLEAST",
                "package_version": "5.2.0",
                "package_name": "Pillow"
            }
        ]
    }
]
```

● 图像分类模型配置文件示例

```
{
    "model_type": "TensorFlow",
    "model_algorithm": "image_classification",
    "metrics": [
        "f1": 0.345294,
        "accuracy": 0.462963,
        "precision": 0.338977,
        "recall": 0.351852
    ],
    "apis": [
        {
            "protocol": "http",
            "url": "/",
            "method": "post",
            "request": {
                "Content-type": "multipart/form-data",
                "data": [
                    {
                        "type": "object",
                        "properties": {
                            "image": {
                                "type": "string"
                            }
                        }
                    }
                ]
            },
            "response": {
                "Content-type": "multipart/form-data",
                "data": [

```

```
        "type": "object",
        "required": [
            "predicted_label",
            "scores"
        ],
        "properties": {
            "predicted_label": {
                "type": "string"
            },
            "scores": {
                "type": "array",
                "items": [
                    {
                        "type": "array",
                        "minItems": 2,
                        "maxItems": 2,
                        "items": [
                            {
                                "type": "string"
                            },
                            {
                                "type": "number"
                            }
                        ]
                    }
                ]
            }
        }
    ],
    "dependencies": [
        {
            "installer": "pip",
            "packages": [
                {
                    "restraint": "ATLEAST",
                    "package_version": "1.15.0",
                    "package_name": "numpy"
                },
                {
                    "restraint": "",
                    "package_version": "",
                    "package_name": "h5py"
                },
                {
                    "restraint": "ATLEAST",
                    "package_version": "1.8.0",
                    "package_name": "tensorflow"
                },
                {
                    "restraint": "ATLEAST",
                    "package_version": "5.2.0",
                    "package_name": "Pillow"
                }
            ]
        }
    ]
}
```

● 自定义镜像类型的模型配置文件示例

```
{
    "model_algorithm": "image_classification",
    "model_type": "Image",
    "swr_location": "100.125.5.235:20202/w00257379-sample/cus-image:test",
    "metrics": {
        "f1": 0.345294,
        "accuracy": 0.462963,
        "precision": 0.338977,
        "recall": 0.351852
    },
    "apis": [
        {
            "protocol": "http",
            "url": "/",
            "method": "post",
            "request": {

```

```
    "Content-type": "multipart/form-data",
    "data": {
        "type": "object",
        "properties": {
            "image": {
                "type": "string"
            }
        }
    },
    "response": {
        "Content-type": "multipart/form-data",
        "data": {
            "type": "object",
            "required": [
                "predicted_label",
                "scores"
            ],
            "properties": {
                "predicted_label": {
                    "type": "string"
                },
                "scores": {
                    "type": "array",
                    "items": [
                        {
                            "type": "array",
                            "minItems": 2,
                            "maxItems": 2,
                            "items": [
                                {
                                    "type": "string"
                                },
                                {
                                    "type": "number"
                                }
                            ]
                        }
                    ]
                }
            }
        }
    }
}
```

12.3 模型推理代码编写说明

- 所有的自定义Python代码必须继承自`BaseService`类，不同类型的模型父类导入语句如[表12-8](#)所示。

表 12-8 `BaseService` 类导入语句

模型类型	父类	导入语句
TensorFlow	TfServingBaseService	from model_service.tfserving_model_service import TfServing BaseService
MXNet	MXNetBaseService	from mms.model_service.mxnet_model_service import MXNetBaseService
PyTorch	PTervingBaseService	from model_service.pytorch_model_service import PTerving BaseService

模型类型	父类	导入语句
Pyspark	SparkServing BaseService	from model_service.spark_model_service import SparkServingBaseService
Caffe	CaffeBaseSer vice	from model_service.caffe_model_service import CaffeBaseService

2. 可以重写的方法有以下几种。

表 12-9 重写方法

方法名	说明
<code>_init_(self, model_name, model_path)</code>	初始化方法，该方法内加载模型及标签等（pytorch和caffe类型模型必须重写，实现模型加载逻辑）。
<code>_preprocess(self, data)</code>	预处理方法，在推理请求前调用，用于将API接口用户原始请求数据转换为模型期望输入数据。
<code>_inference(self, data)</code>	实际推理请求方法（不建议重写，重写后会覆盖modelarts内置的推理过程，运行自定义的推理逻辑）
<code>_postprocess(self, data)</code>	后处理方法，在推理请求完成后调用，用于将模型输出转换为API接口输出

说明

- 通常，用户可以选择重写preprocess和postprocess方法，以实现数据的API输入的预处理和推理结果输出的后处理。
 - 重写 BaseService继承类的初始化方法init可能导致模型“运行异常”
3. 可以使用的属性为模型所在的本地目录，属性名为“self.model_path”。另外pyspark模型在“customize_service.py”中可以使用“self.spark”获取SparkSession对象。
 4. TensorFlow MnistService示例如下。

- 推理代码

```
from PIL import Image
import numpy as np
from model_service.tfserving_model_service import TfServingBaseService

class mnist_service(TfServingBaseService):

    def _preprocess(self, data):
        preprocessed_data = {}

        for k, v in data.items():
            for file_name, file_content in v.items():
                image1 = Image.open(file_content)
                image1 = np.array(image1, dtype=np.float32)
                image1.resize((1, 784))
                preprocessed_data[k] = image1

    return preprocessed_data
```

```
def _postprocess(self, data):
    infer_output = {}
    for output_name, result in data.items():
        infer_output["mnist_result"] = result[0].index(max(result[0]))
    return infer_output
```

- 请求
curl -X POST \ 在线服务地址 \ -F images=@test.jpg
- 返回
{"mnist_result": 7}

在上面的代码示例中，完成了将用户表单输入的图片的大小调整，转换为可以适配模型输入的shape，我们首先通过Pillow库读取“32×32”的图片，调整图片大小为“1×784”以匹配模型输入。在后续处理中，转换模型输出为列表，用于Restful接口输出展示。

13 将 DLS 数据迁移至 ModelArts

由于DLS服务将于2019年5月30日下线，建议您在ModelArts服务中使用深度学习相关的功能。为保证用户的业务数据正常运行，建议您参考如下操作指导完成数据迁移。

代码和数据集（不需迁移）

DLS服务中，用户的代码和数据保存在OBS桶里的，不需要迁移。建议您保存好代码和数据集所在OBS路径，以便业务迁移至ModelArts服务时直接使用。

迁移训练作业参数

DLS服务中，您创建的训练作业可以保存作业参数，这部分数据需要您进行手动迁移，具体操作指导如下：

1. 进入DLS服务管理控制台，在左侧菜单中选择“作业参数管理”，进入作业参数管理界面。
2. 作业参数管理界面，罗列此用户下所有的作业参数。选择需要进行迁移的作业参数，单击“参数配置名称”左侧箭头，查看其配置详情。

图 13-1 查看作业参数的详情

参数配置名称	类型	引擎类型	创建时间	描述
fhfhjgfjh	训练	TensorFlow , TF-1.4.0-python2.7	2019/04/29 12:34:15 GMT+08:00	-

3. 进入ModelArts服务管理控制台，在左侧菜单栏中选择“训练作业”，进入训练作业管理页面。

图 13-2 在 ModelArts 服务中创建训练作业



4. 在训练作业管理页面，单击“创建”。

请根据步骤2中获取的参数配置信息，在ModelArts服务中创建训练作业，并填写对应参数。

说明

由于训练作业将产生费用，建议您创建作业完成后，立即停止或删除训练作业。此操作不影响作业参数的保存。停止和删除训练作业的操作请参见步骤7。

表 13-1 ModelArts 训练作业参数填写说明

类别	参数	填写说明
基本信息 如图图 13-3所示	计费模式	当前只支持按需计费，保持默认即可。
	名称	指定训练作业名称。
	版本	版本信息自动生成，不需设置。
	描述	训练作业描述信息。可不填写。
作业参数 和资源 如图13-4 所示	数据来源	选择“数据存储位置”，并单击“选择”，选择DLS服务中作业参数指定的数据存储路径，即DLS作业参数中“训练数据集”对应的OBS路径。
	算法来源	选择“常用框架”，然后设置如下三个参数信息。 <ul style="list-style-type: none">● AI引擎：选择训练作业使用的引擎类型和对应版本，与DLS作业参数保持一致。如TensorFlow、MXNet。● 代码目录：指定代码目录，即DLS作业参数中“代码目录”对应的OBS路径。● 启动文件：指定启动文件，即DLS作业参数中“启动文件”对应的OBS路径及文件名称。
	运行参数	单击“+”，增加运行参数。参数配置请与DLS作业参数的设置保持一致。
	训练输出位置	选择作业训练输出位置。即DLS作业参数中“训练输出文件路径”对应的OBS路径。
	作业日志路径	选择作业日志存储路径。即DLS作业参数中“作业日志路径”对应的OBS路径。

类别	参数	填写说明
订阅消息和是否保存作业参数 如图13-5所示	资源池	选择一个资源池，建议与DLS作业参数中的“计算节点规格”保持一致。
	计算节点个数	建议与DLS作业参数中的“计算节点个数”保持一致。
	订阅消息	默认关闭，不需设置。
	保存作业参数	勾选保存作业参数选项。
	作业参数名称	设置作业参数名称，建议与DLS参数作业保持一致。
	作业参数描述	设置作业参数描述，建议与DLS参数作业保持一致。

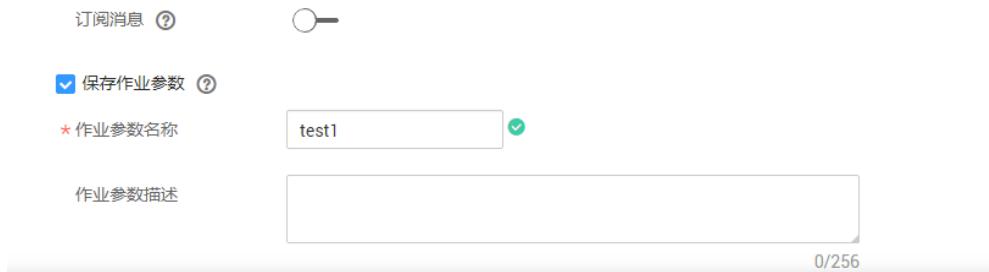
图 13-3 基本信息设置

The screenshot shows the basic information setup process. Step 1: Service Selection (Service Type: Pay-as-you-go, Name: trainjob-9132, Version: V0001, Description: empty). Step 2: Specification Confirmation (Step 2 of 3). Step 3: Completion.

图 13-4 作业参数和资源选择

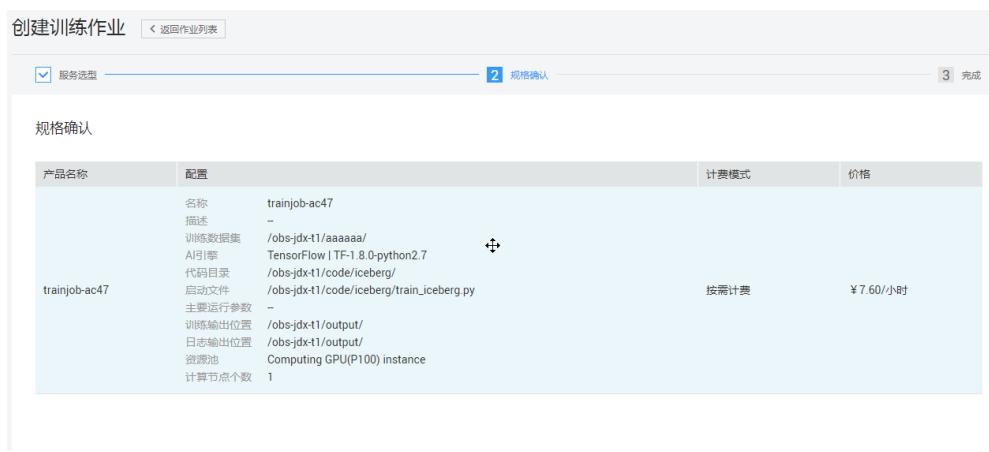
The screenshot shows the job parameter and resource selection interface. It includes sections for Algorithm Source (TensorFlow, Python 2.7), Runtime Parameters (增加运行参数, Add runtime parameters), Training Output Path (/obs-jdx-t1/output), Log Path (/obs-jdx-t1/output), Resource Pool (Computing GPU(P100) instance), and Node Count (1).

图 13-5 订阅消息和是否保存作业参数



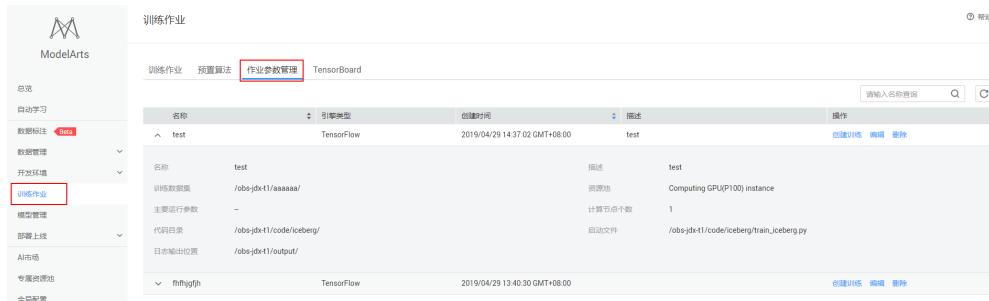
5. 参数填写完成后，单击“下一步”。
6. 在“规格确认”页面，完成信息确认，然后单击“立即创建”，创建训练作业创建并保存作业参数。

图 13-6 确认规格



7. 进入ModelArts“训练作业”管理页面，“删除”或“停止”此新建的训练作业。由于训练作业将产生费用，如果您不需要基于此作业参数进行训练，建议您“停止”或“删除”此训练作业。不管训练作业删除或停止，其对应的作业参数已保存至“作业参数管理”页面。
8. 在ModelArts“训练作业”管理页面，单击“作业参数管理”页签，可查看前面步骤中保存的作业参数。至此，完成DLS服务中一个训练作业参数的迁移。如有多个训练作业参数需要迁移，前参考步骤1至步骤8完成迁移。

图 13-7 查看迁移后的作业参数



A 修订记录

发布日期	修改说明
2019-08-08	<ul style="list-style-type: none">新增提供模型转换功能。 转换模型在线服务的部署增加自动停止功能，并在详情页面增加事件的信息。 部署为在线服务 查看服务详情新增模型二次调优功能。 模型二次调优
2019-07-18	修改Notebook中选择AI引擎的方式。修改如下章节： Notebook简介 创建并打开Notebook
2019-06-30	新增Notebook上传大文件的功能。 使用Notebook上传大文件
2019-06-20	新增关于模型模板的输入输出模式说明。 <ul style="list-style-type: none">预置物体检测模式预置预测分析模式未定义模式
2019-06-03	导入模板时，新增从模板中导入的功能。 <ul style="list-style-type: none">新增：模型模板修改：导入模型 开发环境中，增加“Multi-Engine(Recommend)”引擎。 <ul style="list-style-type: none">修改：Notebook简介、创建并打开Notebook

发布日期	修改说明
2019-05-31	<p>将用户指南拆分成三本文档，分别为自动学习用户指南、AI初学者用户指南、AI工程师用户指南。</p> <p>本文档命名为AI工程师用户指南，除自动学习之外，包含了ModelArts管理控制台的功能操作指导。</p> <p>本次刷新，修改了所有大纲，并优化了每个章节的描述语言。</p>
2019-04-16	<p>修改</p> <ul style="list-style-type: none">● 准备工作章节内容。● 开发环境章节内容。
2019-04-12	<p>新增</p> <ul style="list-style-type: none">● 图像分类简介章节内容。● 物体检测简介章节内容。● 数据标注-文本标注章节内容。● 数据标注-语音内容章节内容。 <p>修改</p> <ul style="list-style-type: none">● 修改了模型管理章节内容。● 图像分类-模型训练章节内容。● 物体检测-模型训练章节内容。● 优化了开发环境章节内容。● 优化了在线服务章节内容。
2019-04-04	<p>新增</p> <ul style="list-style-type: none">● 数据标注-声音分类章节内容。
2019-04-01	<p>修改</p> <ul style="list-style-type: none">● 文档导读。● 准备工作。● 自动学习。● 数据标注-Beta。● 训练作业。● 模型管理。● 部署上线。● AI市场。
2019-03-25	<p>新增</p> <ul style="list-style-type: none">● 新增数据标注-Beta章节内容。
2019-03-18	<p>修改</p> <ul style="list-style-type: none">● 修改了图像分类、物体检测和预测分析章节内容。● 调整专属资源池位置。

发布日期	修改说明
2019-02-22	<p>新增</p> <ul style="list-style-type: none">新增了自动学习声音分类，包括构建声音分类模型流程、声音分类。
2019-01-21	<p>新增</p> <ul style="list-style-type: none">新增了模型包规范。新增Notebook中使用Sync OBS功能介绍、调用ModelArts SDK等内容。
2018-12-21	<p>修改</p> <ul style="list-style-type: none">修改了导入模型中导入模型内容。修改了在线服务、批量服务、边缘服务章节内容。修改了AI市场章节内容，增加了共享API描述。
2018-12-03	<p>修改</p> <ul style="list-style-type: none">调整第三章节结构、修改部分内容。
2018-11-15	<p>修改</p> <ul style="list-style-type: none">修改了使用预置算法快速生成模型快速入门样例数据下载路径及步骤。
2018-11-08	第一次正式发布。