ModelArts

AI 工程师用户指南

文档版本 01

发布日期 2023-10-30

华为技术有限公司

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暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

安全声明

漏洞声明

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该政策可参考华为公司官方网站的网址：https://www.huawei.com/cn/psirt/vul-response-process。
如企业客户须获取漏洞信息，请访问：https://securitybulletin.huawei.com/enterprise/cn/security-advisory。

目 录

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**快速入门**和**最佳实践**。

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，从准备数据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

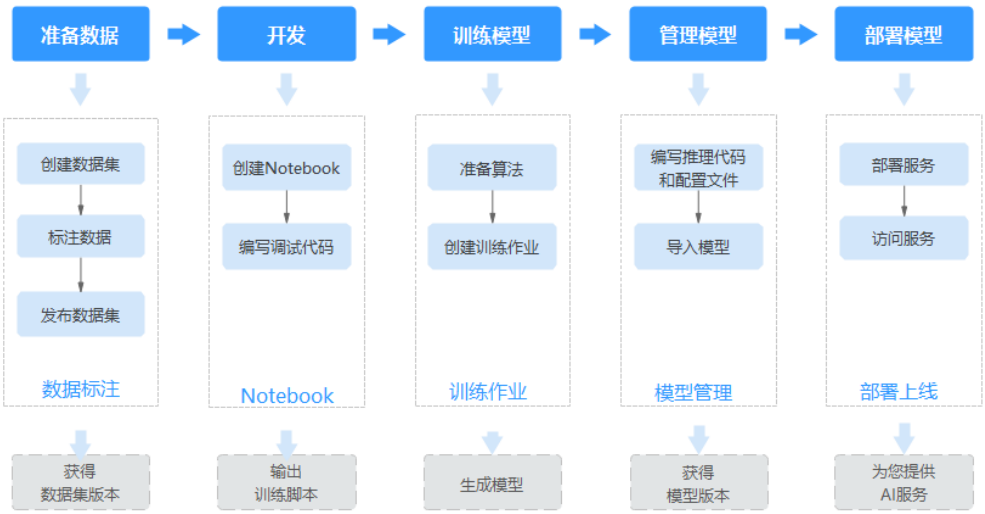


表 1-1 使用流程说明

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Notebook实例
	编写调试代码	在已有的Notebook中编写代码直接构建模型。	JupyterLab简介及常用操作 使用本地IDE开发模型
训练模型	选择算法	创建训练作业前需要先选择算法，可以订阅ModelArts预置的算法，也可以使用自己的算法。	选择算法
	创建训练作业	创建一个训练作业，选择可用的数据集版本，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
管理AI应用	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
	创建AI应用	将训练完成的模型导入至ModelArts创建为AI应用，方便将AI应用部署上线。	创建AI应用
部署AI应用	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。	• 部署为在线服务 • 部署为批量服务 • 部署为边缘服务
	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。	• 访问在线服务 • 查看批量服务预测结果 • 访问边缘服务

使用订阅算法构建模型

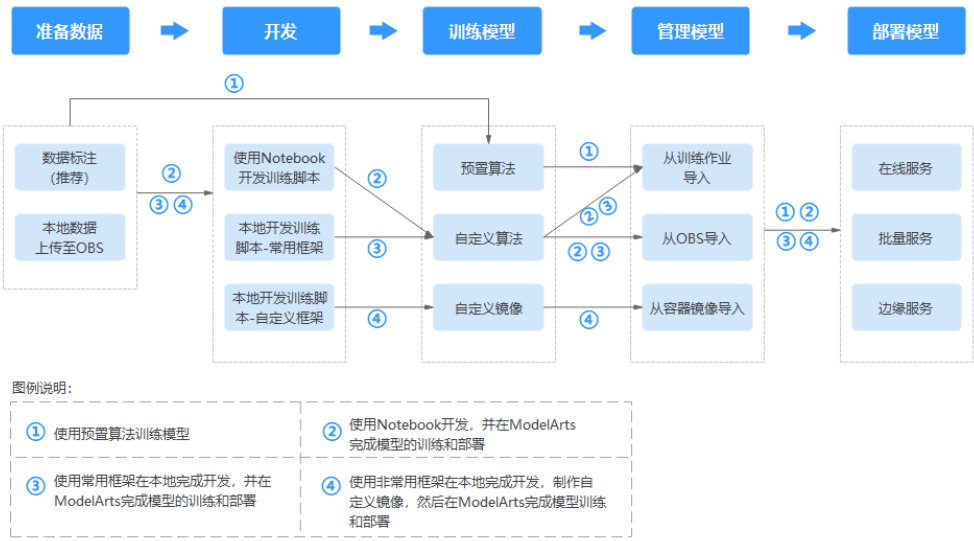
面向有一定AI基础的AI初学者，您可以不关注模型开发，使用自己的业务数据，在AIGallery上订阅ModelArts提供的算法，进行模型训练，从而得到新模型。

图 1-2 AI 初学者使用流程

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，其使用是相对自由的。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 1-3 使用流程详解



2 数据管理（旧版即将下线）

2.1 数据管理简介

说明

当前ModelArts同时存在新版数据集和旧版数据集。本文档主要介绍旧版数据集及数据管理相关功能。

旧版数据集即将下线，推荐用户使用新版数据集及相关功能，具体请参考**数据准备与分析**、**数据标注**、**数据处理**。

新版数据集在旧版的基础上将创建数据集和创建标注任务进行了解耦，创建数据集和创建标注作业分别是独立的任务，使用更灵活。

旧版数据集需要在创建数据集时创建标注任务，不支持分开单独创建数据集和数据标注任务。

在ModelArts中，您可以在“数据管理”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。

数据集的类型

当前ModelArts支持如下类型的数据集。包含图片、音频、文本、表格、视频和其他类别。

- 图片
  - 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。
  - 图像分割：识别出图片中每个物体的轮廓。
- 音频
  - 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
  - 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。
  - 文本三元组：针对文本中的实体片段和实体之间的关系进行标注。

- 表格
  - 表格：适合表格等结构化数据处理。文件格式支持csv。不支持标注，支持对部分表格数据进行预览，但是最多支持100条数据预览。
- 视频
  - 视频标注：识别出视频中每个物体的位置及分类。目前仅支持mp4格式。
- 其他
  - 自由格式：管理的数据可以为任意格式，目前不支持标注，适用于无需标注或开发者自行定义标注的场景。如果您的数据集需存在多种格式数据，或者您的数据格式不符合其他类型数据集时，可选择自由格式的数据集。

图 2-1 自由格式数据集示例

File	Size	Format
image003_1594105979724.jpeg	90.41 KB	jpeg
image005_1594105981161.gif	341.28 KB	gif
iris_1592881356613.csv	2.36 KB	csv
train_1592881356925.csv	8.26 KB	csv
vedio_1594106046159.MP4	430.71 KB	mp4

规格限制

- 除图片类型之外的数据集（如视频、文本、音频等），单个样本大小限制：5GB
- 针对图片类数据集（物体检测、图像分类、图像分割），单个图片大小限制：25MB
- 单个manifest文件大小限制：5GB
- 文本文件单行大小限制：100KB
- 数据管理标注结果文件大小限制：100MB

数据集管理流程及功能简介

图 2-2 标注管理全流程

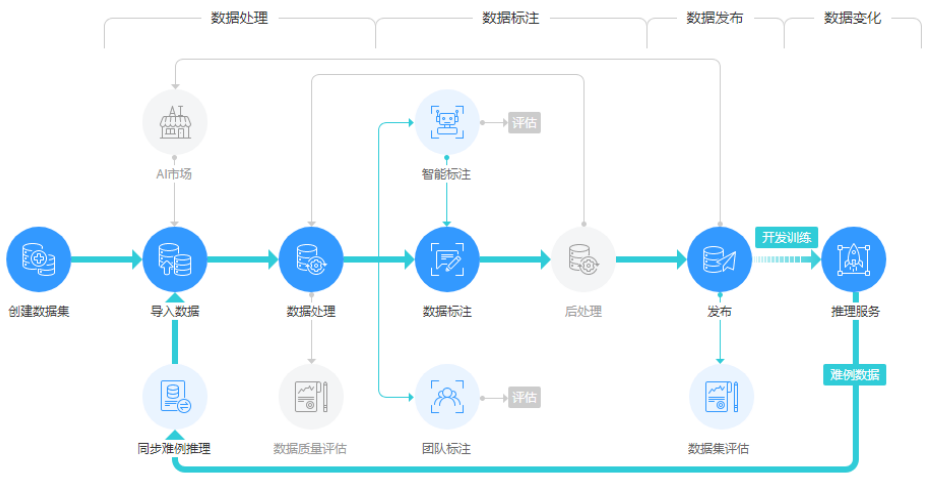


表 2-1 功能介绍

功能	说明
创建数据集（旧版）	创建一个新的数据集。
图像分类 物体检测 文本分类 命名实体 文本三元组 声音分类 语音内容 语音分割 视频标注	针对不同类型的数据集，对数据进行标注。“自由格式”和“表格”类型的数据集暂不支持数据标注。
导入操作	将数据导入数据集中。
导出数据	支持将部分数据导出为新的数据集或者将数据导出至OBS。同时支持对任务历史进行查看和管理。
修改数据集	修改数据集的基本信息。如数据集名称、描述或标签等信息。
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。
智能标注	支持对未标注的数据快速完成数据标注，为您节省70%以上的标注时间。
自动分组	您可以针对您选中的数据，执行自动分组，提升您的数据标注效率。
数据特征	对数据进行特征分析，帮助您了解数据。
团队标注简介	支持多人标注同一个数据集，且支持数据集创建者统一管理标注任务。添加团队及其成员，参与到数据集的标注工作。
数据处理	为了保障数据质量，以免对后续操作（如数据标注、模型训练等）带来负面影响，开发过程通常需要进行数据处理。常见的数据处理类型有：数据校验、数据清洗、数据选择、数据增强。
删除数据集	删除数据集以释放资源。

不同类型数据集支持的功能列表

其中，不同类型的数据集，支持不同的功能，详细信息请参见表2-2。

表 2-2 不同类型的数据集支持的功能

数据集类型	创建数据集	导入数据	导出数据	发布数据集	修改数据集	管理版本	智能标注	团队标注	自动分组	数据特征	一键模型上线
图像分类	支持	支持	支持	支持	支持	支持	支持	支持	支持	支持	支持
物体检测	支持	支持	支持	支持	支持	支持	支持	支持	支持	支持	支持
图像分割	支持	支持	支持	支持	支持	支持	-	-	支持	-	-
声音分类	支持	支持	-	支持	支持	支持	-	-	-	-	-
语音内容	支持	支持	-	支持	支持	支持	-	-	-	-	-
语音分割	支持	支持	-	支持	支持	支持	-	支持	-	-	-
文本分类	支持	支持	-	支持	支持	支持	-	支持	-	-	-
命名实体	支持	支持	-	支持	支持	支持	-	支持	-	-	-
文本三元组	支持	支持	-	支持	支持	支持	-	支持	-	-	-
表格	支持	支持	-	支持	支持	支持	-	-	-	-	-
视频	支持	支持	-	支持	支持	支持	-	-	-	-	-
自由格式	支持	-	支持	支持	支持	支持	-	-	-	-	-

2.2 创建数据集（旧版）

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，都是基于您创建的数据集。

说明

当前ModelArts同时存在新版数据集和旧版数据集。

新版数据集在旧版的基础上将创建数据集和创建标注任务进行了解耦，创建数据集和创建标注作业分别是独立的任务，使用更灵活。

旧版数据集需要在创建数据集时创建标注任务，不支持分开单独创建数据集和数据标注任务。

本文档主要介绍旧版数据集创建流程。新版数据集创建，请参考**创建数据集（New）**。

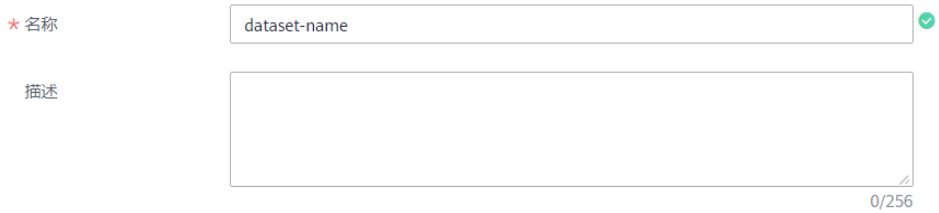
前提条件

- 数据管理功能需要获取访问OBS权限，在未进行委托授权之前，无法使用此功能。在使用数据管理功能之前，请前往“全局配置”页面，使用委托完成访问授权。
- 已创建用于存储数据的OBS桶及文件夹。并且，数据存储的OBS桶与ModelArts在同一区域。
- 需要使用的数据已上传至OBS。详细指导请参见**如何上传数据至OBS**。

操作步骤

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，进入“数据集”管理页面。
2. 单击“创建数据集”，进入“创建数据集”页面，根据数据类型以及数据标注要求，选择创建不同类型的数据集。
   a. 填写数据集基本信息，数据集的“名称”和“描述”。

   图 2-3 数据集基本信息

   b. 根据您的需求，选择“标注场景”和“标注类型”，ModelArts目前支持的类型及其说明请参见**数据集的类型**。

   图 2-4 选择标注场景和标注类型

   c. 针对不同类型的数据集，需填写参数不同，请参考如下类型数据集对应的参数介绍。
      - **图片（图像分类、物体检测、图像分割）**
      - **音频（声音分类、语音内容、语音分割）**
      - **文本（文本分类、命名实体、文本三元组）**
      - **表格**
      - **视频**

- 其他（自由格式）

d. 参数填写无误后，单击页面右下角“创建”。

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、版本管理、修改、导入和删除等操作。不同类型数据集，支持的操作请参见不同类型数据集支持的功能列表

图片（图像分类、物体检测、图像分割）

图 2-5 图像分类和物体检测类型的参数

表 2-3 数据集的详细参数

参数名称	说明
数据集输入位置	选择数据集输入位置的OBS路径。 **说明** 创建数据集时，此OBS路径下的数据会导入数据集，后续若直接在OBS中修改数据，会造成数据集的数据与OBS的数据不一致，可能导致部分数据不可用。
数据集输出位置	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，且不能是“数据集输入位置”的子目录。“数据集输出位置”最好选择一个空目录。

参数名称	说明
添加标签集	• **设置标签名称**：在标签名称文本框中，输入标签名称。标签名称只能是中文、字母、数字、下划线或中划线组成的合法字符串。长度为1～32字符。 • **添加标签**：单击“添加标签”可增加多个标签。 • **设置标签颜色**：仅“物体检测”类型数据集需设置此参数。在每个标签右侧的标签颜色区域下，可在色板中选择颜色，或者直接输入十六进制颜色码进行设置。 • **设置标签属性**：针对“物体检测”类型数据集，在设置完标签颜色后，可在右侧单击加号，增加对应的标签属性。标签属性用于区分同一标签物体的不同属性。例如，黄色小猫、黑色小猫。标签为cat，颜色为不同的标签属性。
启用团队标注	选择是否启用团队标注。图像分割暂不支持团队标注，当选择图像分割类型时，界面不显示此参数。 启用团队标注功能，需填写对应的团队标注任务“名称”、“类型”，同时选择对应的“标注团队”及参与标注的“团队成员”。参数详细介绍请参见创建团队标注任务。 在启用“团队标注”前，需确保您已经在“标注团队”管理页面，添加相应的团队以及成员。如果没有标注团队，可直接从界面链接跳转至“标注团队”页面，添加您的团队并为其添加成员。详细指导请参见团队标注简介。 启用团队标注功能的数据集，在创建完成后，可以在“标注类型”中看到“团队标注”的标识。

音频（声音分类、语音内容、语音分割）

图 2-6 声音分类、语音内容、语音分割类型数据集的参数

参数名称	说明
数据集输入位置	选择数据集输入位置的OBS路径。
数据集输出位置	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，且不能是“数据集输入位置”的子目录。“数据集输出位置”最好选择一个空目录。
添加标签集（声音分类）	仅“声音分类”类型的数据集需设置标签。 ● **设置标签名称**：在标签名称文本框中，输入标签名称。标签名称只能是中文、字母、数字、下划线或中划线组成的合法字符串。长度为1～32字符。 ● **添加标签**：单击“添加标签”可增加多个标签。

参数名称	说明
标签管理（语音分割）	仅“语音分割”类型的数据集，支持多种标签。 ● **单标签** 单标签适用于一段音频标注只有一种类别的音频，通常标注一个标签。 – **设置标签名称**：在“标签名”列输入标签名称。标签名称只能是中文、字母、数字、下划线或中划线组成的合法字符串。长度为1～32字符。 – **设置标签颜色**：在“标签颜色”列设置标签颜色。可在色板中选择颜色，或者直接输入十六进制颜色码进行设置。 ● **多标签** 多标签适用于多维度标注，例如在一段音频标注噪音与人说话的声音两种类别，其中说话的声音还可以标注为不同人的声音。单击“新建标签类别”可添加多个标签类别，一个标签类别可以包含多个标签。“标签类别”和“标签名”只能是中文、字母、数字、下划线或中划线组成的合法字符串。长度为1～32字符。 – **设置标签类别**：在“标签类别”输入标签类别的名称。 – **设置标签名称**：在“标签名”输入标签名称。 – **添加标签**：单击“添加标签”可增加多个标签。
启用语音内容标注（语音分割）	仅“语音分割”类型数据集支持设置，默认关闭。如果启用此功能，支持针对语音内容进行标注。
启用团队标注	仅“语音分割”类型支持团队标注，因此选择创建语音分割类型时，支持设置是否启用团队标注。 启用团队标注功能，需填写对应的团队标注任务“名称”、“类型”，同时选择对应的“标注团队”及参与标注的“团队成员”。参数详细介绍请参见创建团队标注任务。 在启用“团队标注”前，需确保您已经在“标注团队”管理页面，添加相应的团队以及成员。如果没有标注团队，可直接从界面链接跳转至“标注团队”页面，添加您的团队并为其添加成员。详细指导请参见团队标注简介。 启用团队标注功能的数据集，在创建完成后，可以在“标注类型”中看到“团队标注”的标识。

文本（文本分类、命名实体、文本三元组）

图 2-7 文本分类、命名实体、文本三元组类型数据集的参数

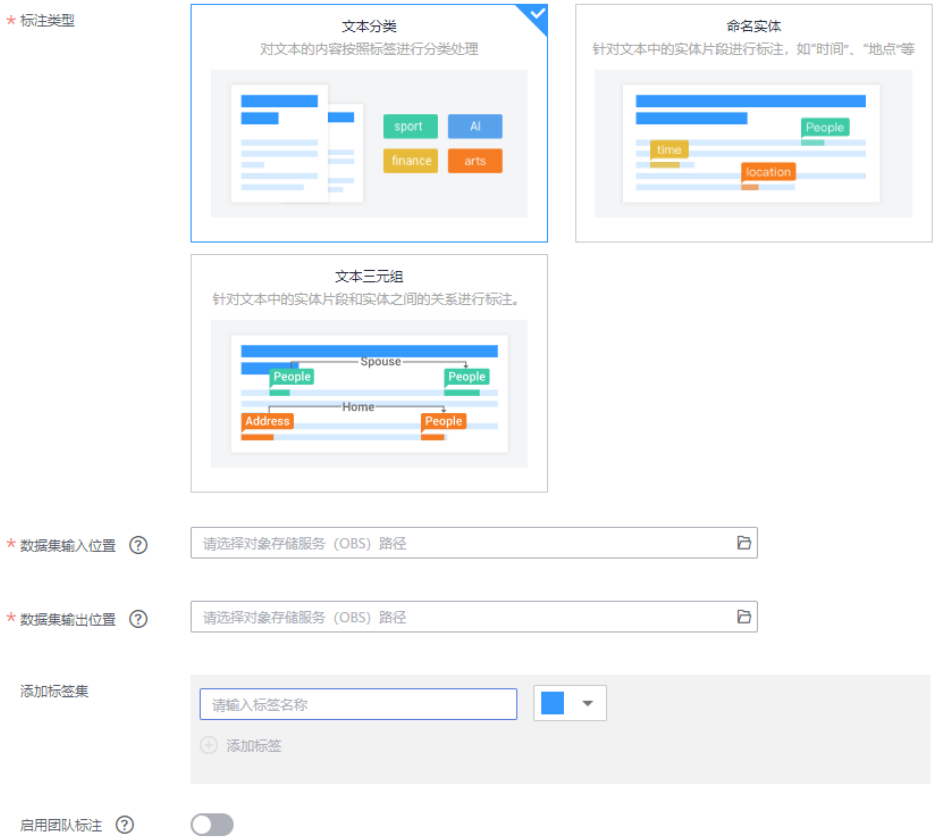
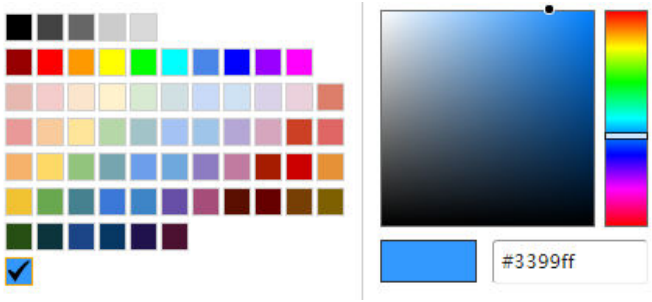
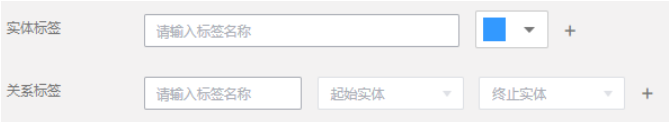


表 2-4 数据集的详细参数

参数名称	说明
数据集输入位置	选择数据集输入位置的OBS路径。 **说明** 文本分类数据只能在执行导入数据操作时识别已标注数据，在此处创建数据集时不能识别已标注数据，建议此处的OBS设置为一个空目录，创建完数据集后再导入已标注数据，导入操作的数据格式要求请参见**从OBS目录导入的规范说明**。
数据集输出位置	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，且不能是“数据集输入位置”的子目录。“数据集输出位置”最好选择一个空目录。

参数名称	说明
添加标签集（文本分类、命名实体）	• **设置标签名称**：在标签名称文本框中，输入标签名称。标签名称只能是中文、字母、数字、下划线或中划线组成的合法字符串。长度为1～32字符。 • **添加标签**：单击“添加标签”可增加多个标签。 • **设置标签颜色**：在每个标签右侧的标签颜色区域下，可在色板中选择颜色，或者直接输入十六进制颜色码进行设置。 
添加标签集（文本三元组）	针对“文本三元组”类型的数据集，需要设置实体标签和关系标签。 • **实体标签**：需设置标签名以及标签颜色。可在颜色区域右侧单击加号增加多个标签。 • **关系标签**：关系标签为两个实体之间的关系。需设置起始实体和终止实体，您需要先添加至少2个实体标签后，再添加关系标签。 
启用团队标注	选择是否启用团队标注。 启用团队标注功能，需填写对应的团队标注任务“名称”、“类型”，同时选择对应的“标注团队”及参与标注的“团队成员”。参数详细介绍请参见创建团队标注任务。 在启用“团队标注”前，需确保您已经在“标注团队”管理页面，添加相应的团队以及成员。如果没有标注团队，可直接从界面链接跳转至“标注团队”页面，添加您的团队并为其添加成员。详细指导请参见团队标注简介。 启用团队标注功能的数据集，在创建完成后，可以在“标注类型”中看到“团队标注”的标识。

表格

图 2-8 表格类型的参数

说明

使用CSV文件时，需要注意以下两点：

- 当数据类型选择String时，默认会把双引号内的数据当作一条，所以同一行数据需要保证双引号闭环，否则会导致数据过大，无法显示。
- 当CSV文件的某一行的列数与定义的Schema不同，则会忽略当前行。

表 2-5 数据集的详细参数

参数名称	说明
存储路径	选择表格数据存储路径（OBS路径），此位置会存放由数据源导入的数据。此位置不能和OBS数据源中的文件路径相同或为其子目录。 创建表格数据集后，在存储路径下会自动生成以下4个目录。 ● annotation：版本发布目录，每次发布版本，会在此目录下生成和版本名称相同的子目录。 ● data：数据存放目录，导入的数据会放在此目录。 ● logs：日志存放目录。 ● temp：临时工作目录。
是否导入	如果您在其他云服务上存储了表格数据，可启用此功能，现支持将存储在对象存储服务(OBS)、数据湖探索(DLI)或MapReduce服务(MRS)的数据导入。

参数名称	说明
数据源（“OBS”）	● “文件路径”：单击输入框右侧按钮，可打开当前帐号下的所有OBS桶，请选择需要导入的数据文件所在目录。 ● “导入是否包含表头”：开启表示导入文件包含表头，此时会将导入文件的第一行作为列名，否则会添加默认列名，自动填写在Schema信息中。 OBS的详细功能说明，请参见《**OBS用户指南**》。
数据源（“DWS”）	● “集群名称”：系统自动将当前账号下的DWS集群展现在列表中，您可以在下拉框中选择您所需的DWS集群。 ● “数据库名称”：根据选择的DWS集群，填写数据所在的数据库名称。 ● “表名称”：根据选择的数据库，填写数据所在的表。 ● “用户名”：输入DWS集群管理员用户的用户名。 ● “密码”：输入DWS集群管理员用户的密码。 DWS的详细功能说明，请参见《**DWS用户指南**》。 **说明** 从DWS导入数据，需要借助DLI的功能，如果用户没有访问DLI服务的权限，需根据页面提示创建DLI的委托。
数据源（“DLI”）	● “队列名称”：系统自动将当前帐号下的DLI队列展现在列表中，您可以在下拉框中选择您所需的队列。 ● “数据库名称”：根据选择的队列展现所有的数据库，请在下拉框中选择您所需的数据库。 ● “表名称”：根据选择的数据库展现此数据库中的所有表。请在下拉框中选择您所需的表。 DLI的详细功能说明，请参见《**DLI用户指南**》。
数据源（“MRS”）	● “集群名称”：系统自动将当前帐号下的MRS集群展现在此列表中，但是流式集群不支持导入操作。请在下拉框中选择您所需的集群。 ● “文件路径”：根据选择的集群，输入对应的文件路径，此文件路径为HDFS路径。 ● “导入是否包含表头”：开启表示导入时将表头同时导入。 MRS的详细功能说明，请参见《**MRS用户指南**》。
Schema信息	表格的列名和对应类型，需要跟导入数据的列数保持一致。请根据您导入的数据输入“列名”，同时选择此列的“类型”。其中支持的类型见**表2-6**。 单击“添加Schema信息”，即可增加一行列。创建数据集时必须指定schema，且一旦创建不支持修改。 从OBS数据源导入数据，会自动获取文件路径下csv文件的schema，如果多个csv文件的schema不一致会报错。

表 2-6 Schema 数据类型说明

类型	描述	存储空间	范围
String	字符串	-	-
Short	有符号整数	2字节	-32768-32767
Int	有符号整数	4字节	-2147483648～2147483647
Long	有符号整数	8字节	-9223372036854775808～9223372036854775807
Double	双精度浮点型	8字节	-
Float	单精度浮点型	4字节	-
Byte	有符号整数	1字节	-128-127
Date	日期类型，描述了特定的年月日，格式：yyyy-MM-dd，例如2014-05-29	-	-
Timestamp	时间戳，表示日期和时间。格式：yyyy-MM-dd HH:mm:ss	-	-
Boolean	布尔类型	1字节	TRUE/FALSE

视频

图 2-9 视频类型的参数

表 2-7 数据集的详细参数

参数名称	说明
数据集输入位置	选择数据集输入位置的OBS路径。
数据集输出位置	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，且不能是“数据集输入位置”的子目录。“数据集输出位置”最好选择一个空目录。
添加标签集	• **设置标签名称**：在标签名称文本框中，输入标签名称。标签名称只能是中文、字母、数字、下划线或中划线组成的合法字符串。长度为1～32字符。 • **添加标签**：单击“添加标签”可增加多个标签。 • **设置标签颜色**：在每个标签右侧的标签颜色区域下，可在色板中选择颜色，或者直接输入十六进制颜色码进行设置。

其他（自由格式）

图 2-10 自由格式类型数据集的参数

表 2-8 数据集的详细参数

参数名称	说明
数据集输入位置	选择数据集输入位置的OBS路径。
数据集输出位置	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，且不能是“数据集输入位置”的子目录。“数据集输出位置”最好选择一个空目录。

2.3 标注数据

2.3.1 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。

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见管理数据集版本。
3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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。

- 对于图像分类数据集，同步数据源操作会以同级目录下的同名“txt”文件作为对应图像的标签。
- 对于物体检测、图像分割数据集，则以同级目录下的同名“xml”文件作为对应图像的标签。

为了快速获取OBS桶中最新数据，可在数据集详情页的“全部”或“未标注”页签中，单击“同步数据源”，快速将通过OBS上传的数据添加到数据集中。

筛选数据

在数据集概览页中，单击页面右上角的“开始标注”，进入数据集的详情页面，默认展示数据集中全部数据。在“全部”、“未标注”或“已标注”页签下，您可以在筛选条件区域，添加筛选条件，快速过滤出您想要查看的数据。

支持的筛选条件如下所示，您可以设置一个或多个选项进行筛选。

- 难例集：难例或非难例。
- 标签：您可以选择全部标签，或者基于您指定的标签，选中其中一个或多个。
- 样本创建时间：1个月内、1天内或自定义，如果选择自定义，可以在时间框中指定明确时间范围。

- 文件名或目录：根据文件名称或者文件存储目录筛选。
- 标注人：选择执行标注操作的帐号名称。
- 样本属性：表示自动分组生成的属性。只有启用了自动分组任务后才可使用此筛选条件。
- 数据属性：暂不支持。

图 2-11 筛选条件

标注图片（手工标注）

数据集详情页中，展示了此数据集中“全部”、“未标注”和“已标注”的图片，默认显示“全部”的图片列表。单击图片，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，单击勾选图片左上角的选择框，进入选择模式，表示图片已勾选。可勾选同类别的多个图片，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. 为选中图片添加标签。
   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。

      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。如果已有标签无法满足要求时，可进入修改数据集页面，添加标签。

   b. 查看“选中文件标签”的信息，确认无误后，单击“确认”。此时，选中的图片将被自动移动至“已标注”页签，且在“未标注”和“全部”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 2-12 添加标签

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。图片缩略图下方默认呈现其对应的标签，您也可以勾选图片，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**

  在数据集详情页面，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。

  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，然后单击确定图标完成修改。

  删除标签：在“选中文件标签”区域中，单击操作列的删除图标删除该标签。此操作仅删除选中图片中的标签。

图 2-13 编辑标签

- **基于标签修改**

  在数据集详情页面，单击“已标注”页签，在图片列表右侧，显示全部标签的信息。

  – 修改标签：单击操作列的编辑按钮，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：单击操作列的删除按钮，在弹出的对话框中，选择“仅删除标签”、“删除标签及仅包含此标签的图片（不删除源文件）”或“删除标签及仅包含此标签的图片（同时删除源文件）”，然后单击“确定”。

图 2-14 全部标签的信息

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“全部”或“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。支持JPG、JPEG、PNG、BMP四种格式图片，单张图片大小不能超过5MB，单次上传的图片总大小不能超过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 2-15 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“全部”、“未标注”或“已标注”页面中，依次选中需要删除的图片，或者选择“选择当前页”选中该页面所有图片，然后单击左上角“删除图片”。在弹出的对话框中，根据实际情况选择是否勾选“同时删除源文件”，确认信息无误后，单击“确定”完成图片删除操作。

其中，被选中的图片，其左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。

说明

如果勾选了“同时删除源文件”，删除图片操作将删除对应OBS目录下存储的图片，此操作可能会影响已使用此源文件的其他数据集或数据集版本，有可能导致展示异常或训练/推理异常。删除后，数据将无法恢复，请谨慎操作。

2.3.2 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，画框边缘不可与待标注的物体的边缘轮廓相交，在此基础之上确保边缘和待标注物体间不要留着空隙，避免背景对模型训练造成干扰。

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见**管理数据集版本**。
3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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。

- 对于图像分类数据集，同步数据源操作会以同级目录下的同名“txt”文件作为对应图像的标签。
- 对于物体检测、图像分割数据集，则以同级目录下的同名“xml”文件作为对应图像的标签。

为了快速获取OBS桶中最新数据，可在数据集详情页的“全部”或“未标注”页签中，单击“同步数据源”，快速将通过OBS上传的数据添加到数据集中。

筛选数据

在数据概览页中，默认展示数据集的概览情况。在界面左上方，单击“开始标注”，进入数据集的详细数据页面，默认展示数据集中全部数据。在“全部”、“未标注”或“已标注”页签下，您可以在筛选条件区域，添加筛选条件，快速过滤出您想要查看的数据。

支持的筛选条件如下所示，您可以设置一个或多个选项进行筛选。

- 难例集：难例或非难例。
- 标签：您可以选择全部标签，或者基于您指定的标签，选中其中一个或多个。
- 样本创建时间：1个月内、1天内或自定义，如果选择自定义，可以在时间框中指定明确时间范围。

- 文件名或目录：根据文件名称或者文件存储目录筛选。
- 标注人：选择执行标注操作的帐号名称。
- 样本属性：表示自动分组生成的属性。只有启用了**自动分组**任务后才可使用此筛选条件。
- 数据属性：暂不支持。

图 2-16 筛选条件

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“全部”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。在标注页面，常用按钮的使用可参见**表2-10**。
2. 在页面左侧工具栏选择合适的标注图形，系统默认的标注图形为矩形。本示例使用矩形工具进行标注。

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。

表 2-9 支持的标注框

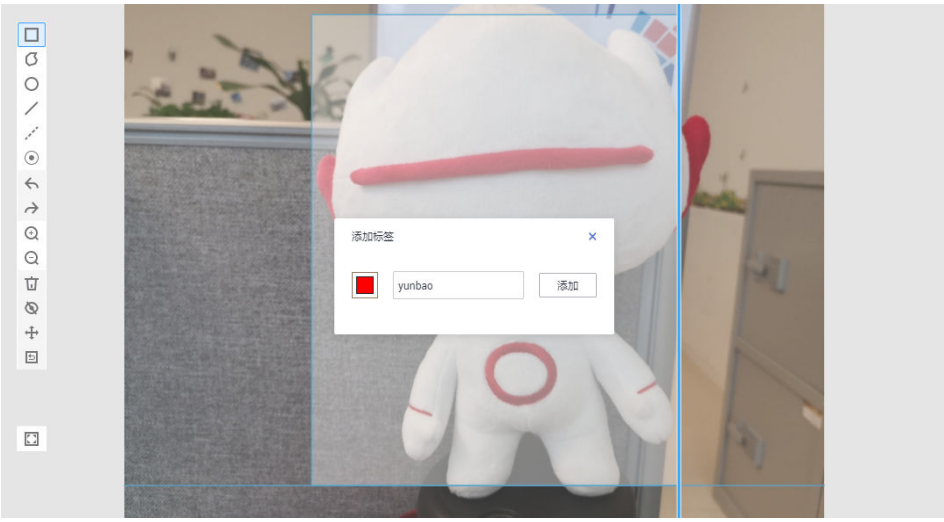
图标	使用说明
□	矩形。鼠标单击标注对象左上角边缘位置，界面将出现矩形框，移动鼠标使得矩形框覆盖标注对象，然后单击完成标注。
⬠	多边形。在标注对象所在范围内，鼠标左键单击完成一个点的标注，沿着物体的形状边缘，通过鼠标指定多个点，最终单击到第一个点的位置，由所有的点组成一个多边形形状。使得需标注的对象在此标注框内。
○	圆形。在标注对象中，选择物体的中心点位置，单击鼠标确定圆心，然后移动鼠标，使得圆形框覆盖标注对象，然后再单击鼠标完成标注。

图标	使用说明
	直线。在标注对象中，选择物体的起始点，单击鼠标确定直线的起始点，然后使得直线覆盖标注对象，然后再单击鼠标完成标注。
	虚线。在标注对象中，选择物体的起始点，单击鼠标确定虚线的起始点，然后使得虚线覆盖标注对象，然后再单击鼠标完成标注。
	点。单击图片中的物体所在位置，即可完成点的标注。

3. 在弹出的添加标签文本框中，直接输入新的标签名，在文本框前面选中标签颜色，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”。

   逐步标注图片中所有物体所在位置，一张图片可添加多个标签。完成一张图片标注后，可单击图片下方图片列表，快速选中其他未标注的图片，然后在标注页面中执行标注操作。

图 2-17 添加物体检测标签



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。

   选中的图片被自动移动至“已标注”页签，且在“未标注”和“全部”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。

表 2-10 标注界面的常用按钮

按钮图标	功能说明
	撤销上一个操作。
	重做上一个操作。

按钮图标	功能说明
	放大图片。
	缩小图片。
	删除当前图片中的所有标注框。
	显示或隐藏标注框。只有在已标注图片中可使用此操作。
	拖动，可将标注好的框拖动至其他位置，也可以选择框的边缘，更改框的大小。
	复位，与上方拖动为同组操作，当执行了拖动后，可以单击复位按钮快速将标注框恢复为拖动前的形状和位置。
	全屏显示标注的图片。

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**

  在数据集详情页面，单击“已标注”页签，然后在图片列表中选中待修改的图片，单击图片跳转到标注页面，在右侧“当前文件标签”区域中对图片信息进行修改。

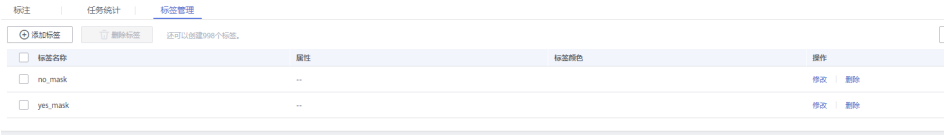
  - 修改标签：“标注”区域中，单击编辑图标，在文本框中输入正确的标签名，然后单击确定图标完成修改。也可以单击标签，在图片标注区域，调整标注框的位置和大小，完成调整后，单击其他标签即可保存修改。
  - 删除标签：在“标注”区域中，单击删除图标即可删除此图片中的标签。

    标签删除后，单击页面左上角的“返回数据标注预览”离开标注页面，在弹出对话框中保存标注。图标的标签全部删除后，该图片会重新回到“未标注”页签。

图 2-18 编辑物体检测标签

- **基于标签修改**

  在数据标注页面，单击“标签管理”页签，即可显示全部标签的信息显示全部标签的信息。

  – 修改标签：单击操作列的“修改”按钮，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“删除”按钮，在弹出的对话框中，根据界面提示选择删除对象，然后单击“确定”。

  说明

  删除后的标签无法再恢复，请谨慎操作。

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“全部”或“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。支持JPG、JPEG、PNG、BMP四种格式图片，单张图片大小不能超过5MB，单次上传的图片总大小不能超过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 2-19 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“全部”、“未标注”或“已标注”页面中，依次选中需要删除的图片，或者选择“选择当前页”选中该页面所有图片，然后单击左上角“删除图片”。在弹出的对话框中，根据实际情况选择是否勾选“同时删除源文件”，确认信息无误后，单击“确定”完成图片删除操作。

其中，被选中的图片，其左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。

说明

如果勾选了“同时删除源文件”，删除图片操作将删除对应OBS目录下存储的图片，此操作可能会影响已使用此源文件的其他数据集或数据集版本，有可能导致展示异常或训练/推理异常。删除后，数据将无法恢复，请谨慎操作。

2.3.3 图像分割

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。您可以通过在ModelArts控制台进行标注，也可以对已标注图片修改或删除标签进行重新标注。

针对图像分割场景，开始标注前，您需要了解：

- 图片中需要提取轮廓的物体都要标注。
- 支持使用多边形标注和极点标注。
  - 多边形标注，根据目标物体的轮廓绘制多边形。
  - 极点标注，在目标物体轮廓的最上、最左、最下、最右的位置分别标注四个极点，极点要在物体的轮廓上。系统将根据标注的极点推理出物体的轮廓。对于背景比较复杂的图片，极点标注效果不佳，推荐使用多边形标注。
- 多边形标注时，标注框或极点，必须在图片范围内，超出图片将导致后续作业异常。

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见管理数据集版本。
3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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。

- 对于图像分类数据集，同步数据源操作会以同级目录下的同名“txt”文件作为对应图像的标签。
- 对于物体检测、图像分割数据集，则以同级目录下的同名“xml”文件作为对应图像的标签。

为了快速获取OBS桶中最新数据，可在数据集详情页的“全部”或“未标注”页签中，单击“同步数据源”，快速将通过OBS上传的数据添加到数据集中。

筛选数据

在数据概览页中，默认展示数据集的概览情况。在界面右上方，单击“开始标注”，进入数据集的详细数据页面，默认展示数据集中全部数据。在“全部”、“未标注”或“已标注”页签下，您可以在筛选条件区域，添加筛选条件，快速过滤出您想要查看的数据。

支持的筛选条件如下所示，您可以设置一个或多个选项进行筛选。

- 难例集：难例或非难例。
- 标签：您可以选择全部标签，或者基于您指定的标签，选中其中一个或多个。

- 样本创建时间：1个月内、1天内或自定义，如果选择自定义，可以在时间框中指定明确时间范围。
- 文件名或目录：根据文件名称或者文件存储目录筛选。
- 标注人：选择执行标注操作的帐号名称。
- 样本属性：表示自动分组生成的属性。只有启用了**自动分组**任务后才可使用此筛选条件。
- 数据属性：暂不支持

图 2-20 筛选条件

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“全部”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。在标注页面，常用按钮的使用可参见**表2-12**。
2. 选择标注方式。

   在标注页面，上方工具栏提供了常用的**标注方式**及**常用按钮**，系统默认的标注方式为多边形标注。选择多边形标注或极点标注。

   说明

   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方式进行标注。

图 2-21 工具栏

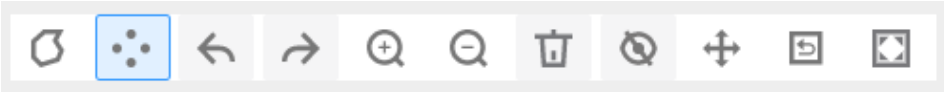


表 2-11 标注方式

图标	使用说明
	多边形。在标注对象所在范围内，鼠标左键单击完成一个点的标注，沿着物体的形状边缘，通过鼠标指定多个点，最终单击到第一个点的位置，由所有的点组成一个多边形形状。使得需标注的对象在此标注框内。

图标	使用说明
	极点标注。在目标物体轮廓的最上、最左、最下、最右的位置分别标注四个极点，极点要在物体的轮廓上。系统将根据标注的极点推理出物体的轮廓。

表 2-12 工具栏常用按钮

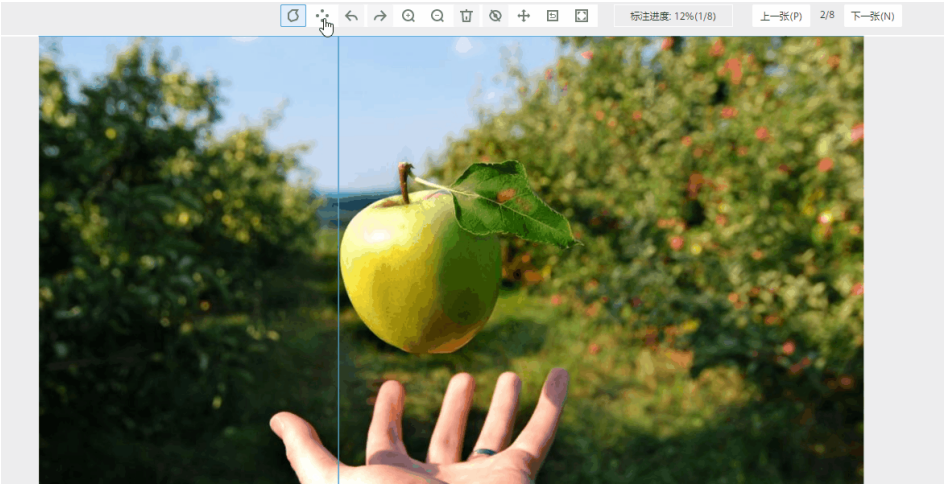
按钮图标	功能说明
	撤销上一个操作。
	重做上一个操作。
	放大图片。
	缩小图片。
	删除当前图片中的所有标注框。
	显示或隐藏标注框。只有在已标注图片中可使用此操作。
	拖动，可将标注好的框拖动至其他位置，也可以选择框的边缘，更改框的大小。
	复位，与上方拖动为同组操作，当执行了拖动后，可以单击复位按钮快速将标注框恢复为拖动前的形状和位置。
	全屏显示标注的图片。

3. 标注物体。

   以极点标注为例。识别图片中的物体，单击左键分别定位物体的最上、最左、最下、最右的位置点。确定位置后，将弹出对话框，填入标签名称，单击确定添加物体的标签。确定后系统将自动推理出物体的轮廓。

   完成一张图片标注后，可单击图片下方图片列表，快速选中其他未标注的图片，然后在标注页面中执行标注操作。

图 2-22 标注物体轮廓



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。

   选中的图片被自动移动至“已标注”页签，且在“未标注”和“全部”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。

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。

修改标注信息

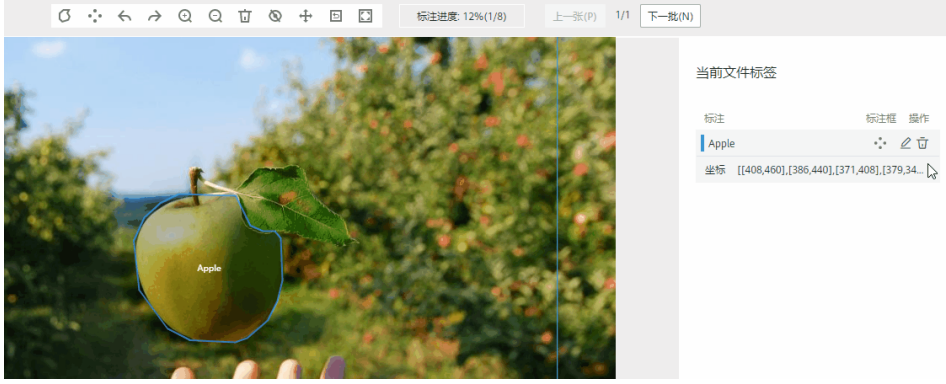
当数据完成标注后，您还可以进入已标注页签，对已标注的数据进行修改。

在数据集详情页面，单击“已标注”页签，然后在图片列表中选中待修改的图片，单击图片跳转到标注页面，在右侧“当前文件标签”区域中单击此图片已添加的标注信息。

- 修改标签：“标注”区域中，单击编辑图标，在弹出框中输入正确的标签名或标签颜色，然后单击“确定”完成修改。也可以单击标签，在图片标注区域，调整标注框的位置和大小，完成调整后，单击其他标签即可保存修改。
- 修改图片标注信息：在图片展示区，显示物体边缘，可单击蓝色圆点，将标注框调整至物体边缘。
- 删除标签：在“标注”区域中，单击删除图标即可删除此图片中的标签。图片的标签全部删除后，该图片会重新回到“未标注”页签。

标注标注信息修改后，单击页面左上角的“返回数据标注预览”离开标注页面，在弹出对话框中单击“确定”保存修改。

图 2-23 编辑标注信息



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“全部”或“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。支持JPG、JPEG、PNG、BMP四种格式图片，单张图片大小不能超过5MB，单次上传的图片总大小不能超过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 2-24 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“全部”、“未标注”或“已标注”页面中，依次选中需要删除的图片，或者选择“选择当前页”选中该页面所有图片，然后单击左上角“删除图片”。在弹出的对话框中，根据实际情况选择是否勾选“同时删除源文件”，确认信息无误后，单击“确定”完成图片删除操作。

其中，被选中的图片，其左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。

说明

如果勾选了“同时删除源文件”，删除图片操作将删除对应OBS目录下存储的图片，此操作可能会影响已使用此源文件的其他数据集或数据集版本，有可能导致展示异常或训练/推理异常。删除后，数据将无法恢复，请谨慎操作。

2.3.4 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。

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见**管理数据集版本**。
3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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. 在“未标注”页签文本列表中，页面左侧罗列“标注对象列表”。在列表中单击需标注的文本对象，选择右侧“标签集”中的标签进行标注。一个标注对象可添加多个标签。

   以此类推，不断选中标注对象，并为其添加标签。

   图 2-25 文本分类标注

2. 当所有的标注对象都已完成标注，单击页面下方“保存当前页”完成“未标注”列表的文本标注。

添加标签

- 在“未标注”页签添加：单击页面中标签集右侧的加号，然后在弹出的“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 2-26 添加标签（1）

- 在“已标注”页签添加：在右侧单击页面中全部标签加号，然后在弹出的“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 2-27 添加标签（2）

图 2-28 新增标签

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。您也可以在右侧的“全部标签”中了解当前数据集支持的所有标签信息。

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。

  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。

- **基于标签修改**

  在数据集详情页，单击“标签管理”页签，标签管理页显示全部标签的信息。

  – 修改：在标签管理页，单击操作列的“修改”，然后在文本框中修改标签名称，选择标签颜色，单击“确定”完成修改。
  – 删除：在标签管理页，单击操作列的“删除”，单击“确定”完成删除。

  说明

  删除后的标签无法恢复，请谨慎操作。

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，根据需上传文件的基本情况，完成设置后选择上传文件。

   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。“文本与标签分割符”与“多标签分割符”不能选同一个。

   – “模式”：选择“文本和标注合并”或“文本和标注分离”模式。界面中已给出示例，请参考示例判断需添加的文件属于哪一种模式。
   – “文本与标签分隔符”：可设置为“Tab键”、“空格”、“分号”、“逗号”或“其他”。选择“其他”时，可以在右侧文本框中输入对应的分隔符。
   – “多标签分隔符”：可设置为“Tab键”、“空格”、“分号”、“逗号”或“其他”。选择“其他”时，可以在右侧文本框中输入对应的分隔符。

图 2-29 添加文件

3. 在添加文件对话框中，单击“上传文件”，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”或“已标注”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对象，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，选中待删除的文本对象，然后单击“删除”，删除单个文本。或者选择“选择当前页”选中该页面所有文本，然后单击左上角“删除”，即可完成当前页所有文本的删除操作。

其中，被选中的文本，其背景将显示为蓝色。

2.3.5 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。

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见管理数据集版本。

3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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. 在“未标注”页签文本列表中，页面左侧罗列“标注对象列表”。在列表中单击需标注的文本对象，在右侧标签集下显示的文本内容中选中需要标注的部分，然后选择右侧“标签集”中的标签进行标注。一个标注对象可添加多个标签。

   以此类推，不断选中标注对象，并为其添加标签。

   图 2-30 命名实体标注

2. 单击页面下方“保存当前页”完成文本标注。

添加标签

- 在“未标注”页签添加：单击页面中标签集右侧的加号，然后在弹出的“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 2-31 添加命名实体标签（1）

- 在“已标注”页签添加：在右侧单击页面中全部标签加号，然后在弹出的“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 2-32 添加命名实体标签（2）

图 2-33 新增命名实体标签

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。您也可以在右侧的“全部标签”中了解当前数据集支持的所有标签信息。

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×删除单个标签。

- **基于标签修改**

  在数据集详情页，单击“标签管理”页签，标签管理页显示全部标签的信息。

  – 修改：在标签管理页，单击操作列的“修改”，然后在文本框中修改标签名称，选择标签颜色，单击“确定”完成修改。
  – 删除：在标签管理页，单击操作列的“删除”，单击“确定”完成删除。

说明

删除后的标签无法恢复，请谨慎操作。

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，根据需上传文件的基本情况，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 2-34 添加文件

3. 在添加文件对话框中，单击“上传文件”，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对象，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，选中待删除的文本对象，然后单击“删除”，删除单个文本。或者选择“选择当前页”选中该页面所有文本，然后单击左上角“删除”，即可完成当前页所有文本的删除操作。

其中，被选中的文本，其背景将显示为蓝色。

2.3.6 文本三元组

三元组标注适用于标注出语句当中形如（主语/Subject，谓词/Predicate，宾语/Object）结构化知识的场景，标注时不但可以标注出语句当中的实体，还可以标注出实体之间的关系，其在依存句法分析、信息抽取等自然语言处理任务中经常用到。

文本三元组类型的数据标注，需要关注两种标签，“实体标签”和“关系标签”。“关系标签”需设置对应的“起始实体”和“终止实体”。

- 支持设置多个“实体标签”和“关系标签”。一个文本数据中，也可以标注多个“实体标签”和“关系标签”

- 创建数据集时定义的“实体标签”，不支持删除。

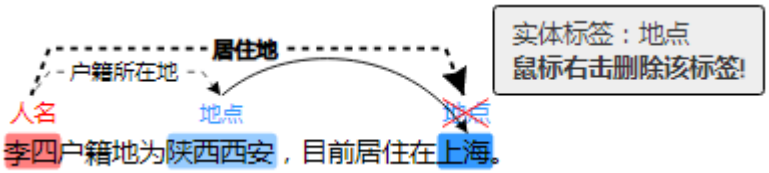
注意事项

在开始标注之前，需确保数据集对应的“实体标签”和“关系标签”已定义好。“关系标签”需设置对应的“起始实体”和“终止实体”。“关系标签”只能添加至其设置好的“起始实体”和“终止实体”之间。

例如，如图2-35所示，当两个文本都被标注为“地点”，那么针对这两个实体，无法添加本示例中的任意一个关系标签。当无法添加某个关系标签时，界面将显示一个红色的叉号，如图2-36所示。

图 2-35 实体标签和关系标签的示例

图 2-36 无法添加关系标签



开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见管理数据集版本。
3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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。

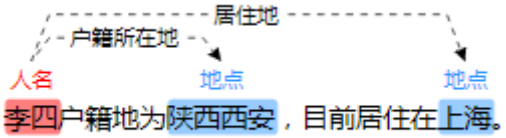
1. 在“未标注”页签文本列表中，页面左侧罗列“标注对象列表”。在列表中单击需标注的文本对象，选中相应文本内容，在页面呈现的实体类型列表中选择实体名称，完成实体标注。

   图 2-37 实体标注

2. 在完成多个实体标注后，鼠标左键依次单击起始实体和终止实体，在呈现的关系类型列表中选择一个对应的关系类型，完成关系标注。

   图 2-38 关系标注

3. 当所有的标注对象都已完成标注，单击页面下方“保存当前页”完成“未标注”列表的文本标注。

说明

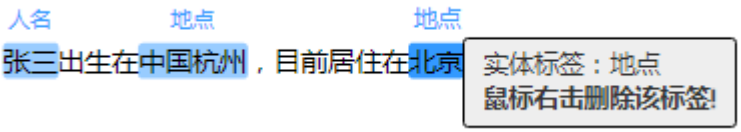
“文本三元组”类型的数据集，不支持在标注页面修改标签，需要进入“修改数据集”页面，修改“实体标签”和“关系标签”。

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。

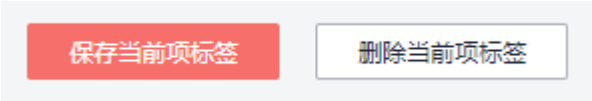
在数据集详情页，单击“已标注”页签，在左侧文本列表中选中一行文本，右侧区域显示具体的标注信息。将鼠标移动至对应的实体标签或关系类型，单击鼠标右键，可删除此标注。单击鼠标左键，依次单击连接起始实体和终止实体，可增加关系类型，增加关系标注。

图 2-39 在文本中修改标签



您也可以在单击页面下方的“删除当前项标签”按钮，删除选中文本对象中的所有标签。

图 2-40 删除当前项标签



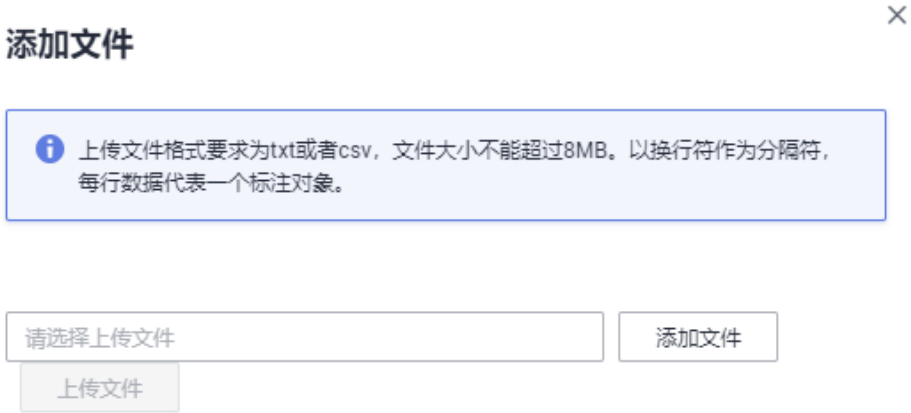
添加文件

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 2-41 添加上传文件



3. 在添加文件对话框中，单击“上传文件”，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“标注对象列表”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，选中待删除的文本，然后单击“删除”，删除单个文本。或者勾选“选择当前页”选中该页面所有文本，然后单击左上角“删除”，即可完成当前页所有文本的删除操作。

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”按钮为灰色，无法执行删除操作。

2.3.7 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。

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见管理数据集版本。
3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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，可在数据集详情页的“未标注”页签中，单击“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。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”下侧的文本框中设置标签。

      方式一（已存在标签）：单击“标签”下方的文本框，在快捷键下拉列表中选择快捷键，然后在标签文本输入框中选择已有的标签名称，然后单击“确定”。

      方式二（新增标签）：在“标签”下方的文本框中，在快捷键下拉列表中选择快捷键，然后在标签文本输入框中输入新的标签名称，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。

说明

快捷键的使用说明：为标签指定快捷键后，当您选择一段音频后，在键盘中按一下快捷键，即可为此音频增加为此快捷键对应的标签。例如“aa”标签对应的快捷键是“1”，在数据标注过程中，选中1个或多个文件，按“1”，界面将提示是否需要将此文件标注为“aa”标签，单击确认即可完成标注。

快捷键对应的是标签，1个标签对应1个快捷键。不同的标签，不能指定为同一个快捷键。快捷键的使用，可以大大提升标注效率。

图 2-42 添加音频标签

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**

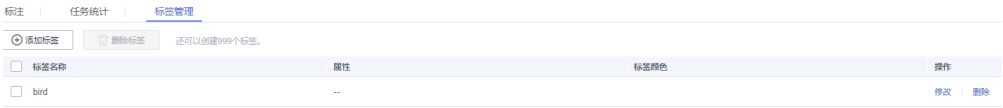
  在数据标注页面，单击“已标注”页签，然后在音频列表中选中待修改的音频（选择一个或多个）。在右侧标签信息区域中对标签进行修改。

  - 修改标签：在“选中文件标签”区域中，单击操作列的编辑图标，然后在文本框中输入正确的标签名，然后单击确定图标完成修改。
  - 删除标签：在“选中文件标签”区域中，单击操作列的删除图标删除该标签。

- **基于标签修改**

  在数据标注页面，单击“标签管理”页签，在标签管理页，显示全部标签的信息。

图 2-43 全部标签的信息



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，选中需要删除的音频，或者选择“选择当前页”选中该页面所有音频，然后单击左上角“删除音频”，在弹出的对话框中，根据实际情况选择是否勾选“同时删除源文件”，确认信息无误后，单击“确定”完成音频删除操作。

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。

说明

如果勾选了“同时删除源文件”，删除音频操作是将删除对应OBS目录下存储的音频。此操作可能会影响已使用此源文件的其他数据集或数据集版本，有可能导致展示异常或训练/推理异常。删除后，数据将无法恢复，请谨慎操作。

2.3.8 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。

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见管理数据集版本。

3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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。

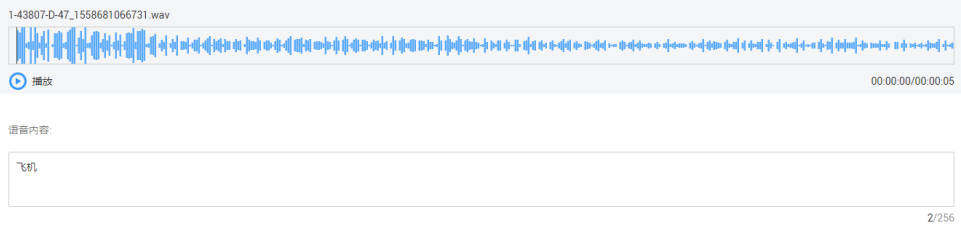
为了快速获取OBS桶中最新数据，可在数据集详情页的“未标注”页签中，单击“同步数据源”，快速将通过OBS上传的数据添加到数据集中。

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，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。

图 2-44 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，选中需要删除的音频，然后单击左上角“删除音频”，在弹出的对话框中，根据实际情况选择是否勾选“同时删除源文件”，确认信息无误后，单击“确定”完成音频删除操作。

说明

如果勾选了“同时删除源文件”，删除音频操作是将删除对应OBS目录下存储的音频。此操作可能会影响已使用此源文件的其他数据集或数据集版本，有可能导致展示异常或训练/推理异常。删除后，数据将无法恢复，请谨慎操作。

2.3.9 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。

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见**管理数据集版本**。
3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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，可在数据集详情页的“未标注”页签中，单击“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”文本框中填写音频标签和内容。

图 2-45 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。

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叉号删除标注，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，选中需要删除的音频，然后单击左上角“删除音频”，在弹出的对话框中，根据实际情况选择是否勾选“同时删除源文件”，确认信息无误后，单击“确定”完成音频删除操作。

说明

如果勾选了“同时删除源文件”，删除音频操作是将删除对应OBS目录下存储的音频。此操作可能会影响已使用此源文件的其他数据集或数据集版本，有可能导致展示异常或训练/推理异常。删除后，数据将无法恢复，请谨慎操作。

2.3.10 视频标注

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

开始标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理> 数据集”，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概览页。

   此操作默认进入数据集当前版本的概览页，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“当前版本。”详细操作指导请参见**管理数据集版本**。
3. 在数据集概览页中，单击右上角“开始标注”，进入数据集详情页。数据集详情页默认展示此数据集下全部数据。

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，可在数据集详情页的“未标注”页签中，单击“同步数据源”，快速将通过OBS上传的数据添加到数据集中。

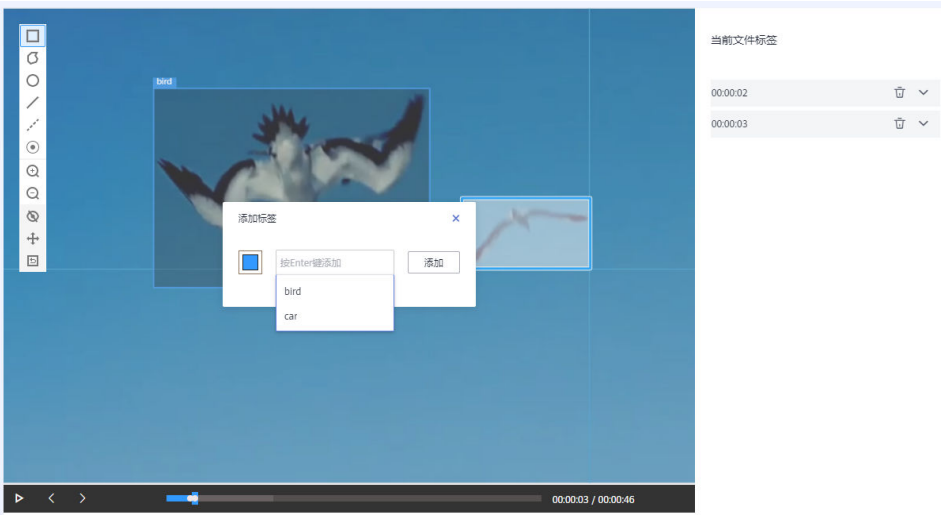
视频标注

数据集详情页中，展示了此数据集中“未标注”和“已标注”的视频。

1. 在“未标注”页签左侧视频列表中，单击目标视频文件，打开标注页面。
2. 在标注页面中，播放视频，当视频播放至待标注时间时，单击进度条中的暂停按钮，将视频暂停至某一画面。
3. 在左侧区域选择标注框，默认为矩形框。使用鼠标在视频画面中框出目标，然后在弹出的添加标签文本框中，直接输入新的标签名，在文本框前面选中标签颜色，单击“添加”完成1个物体的标注。如果已存在标签，从下拉列表中选择已有的标签，然后单击“添加”完成标注。逐步此画面中所有物体所在位置，一张画面可添加多个标签。

   支持的标注框与“物体检测”类型一致，详细描述请参见**物体检测**章节的**表2-9**。

   图 2-46 视频标注



4. 上一个暂停画面标注完成后，在进度条处单击播放按钮继续播放，在需要标注的画面中暂停，然后重复执行步骤**3**完成整个视频的标注。

界面右侧将呈现当前视频带标注的时间点。

图 2-47 当前文件标签信息

5. 单击页面左上角“返回数据标注预览”，页面将自动返回数据集详情页面，同时，标注好的视频将呈现在“已标注”页签下。

修改标注

当数据完成标注后，您还可以进入“已标注”页签，删除标注数据。

- 单击目标编号操作列的 ，删除该段视频的标注。您也可以单击标注视频文件上方的叉号删除标注，然后单击“确认标注”。

在“已标注”页签下，单击目标视频文件，在标注页面右侧的“当前文件标签”下，可单击时间点右侧小三角展开详情，您可以修改或删除标签。

- 修改标签：单击标签右侧的编辑按钮，标签名称可进行修改。
- 删除标签：单击标签右侧的删除按钮，将直接删除此标签。如果单击画面时间右侧的删除按钮，将删除此画面下的所有标签。

图 2-48 修改标注

删除视频

通过数据删除操作，可将需要丢弃的视频数据快速删除。

在“全部”、“未标注”或“已标注”页面中，依次选中需要删除的视频，或者选择“选择当前页”选中该页面所有视频，然后单击左上角“删除视频”。在弹出的对话框中，根据实际情况选择是否勾选“同时删除源文件”，确认信息无误后，单击“确定”完成视频删除操作。

其中，被选中的视频，其左上角将显示为勾选状态。如果当前页面无选中视频时，“删除视频”按钮为灰色，无法执行删除操作。

说明

如果勾选了“同时删除源文件”，删除视频操作将删除对应OBS目录下存储的视频，此操作可能会影响已使用此源文件的其他数据集或数据集版本，有可能导致展示异常或训练/推理异常。删除后，数据将无法恢复，请谨慎操作。

2.4 导入数据

2.4.1 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。
- 确保数据存储的OBS桶与ModelArts在同一区域。

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**从OBS目录导入的规范说明**。当前只有“图像分类”、“物体检测”、“表格”、“文本分类”和“声音分类”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文件的规范请参见**导入Manifest文件的规范说明**。

说明

导入“物体检测”类型数据集前，您需要保证标注文件的标注范围不超过原始图片大小，否则可能存在导入失败的情况。

表 2-13 不同类型数据集支持的导入方式

数据集类型	OBS目录导入	Manifest文件导入
图像分类	支持 格式规范：**图像分类**	支持 格式规范：**图像分类**

数据集类型	OBS目录导入	Manifest文件导入
物体检测	支持 格式规范：物体检测	支持 格式规范：物体检测
图像分割	支持 格式规范：图像分割	支持 格式规范：图像分割
声音分类	支持 格式规范：声音分类	支持 格式规范：声音分类
语音内容	-	支持 格式规范：语音分割
语音分割	-	支持 格式规范：语音内容
文本分类	支持 格式规范：文本分类	支持 格式规范：文本分类
命名实体	-	支持 格式规范：文本命名实体
文本三元组	-	支持 格式规范：文本三元组
表格	支持 还支持从DWS、DLI、MRS导入数据。 格式规范：表格	-
视频	-	支持 格式规范：视频标注
自由格式	-	-

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“数据管理 >数据集”，进入“数据集”管理页面。
2. 在数据集所在行，单击操作列的“更多 > 导入”。

   或者，您可以单击数据集名称，进入数据集“概览”页，在页面右上角单击“导入”。
3. 在“导入”对话框中，设置“导入方式”为“OBS目录”，然后在“对象存储服务（OBS）目录”中，设置数据存储的路径。然后单击“确定”。

说明

针对“表格”类型的数据集，导入时，支持从OBS、DWS、DLI和MRS等数据源。其导入时的设置和数据要求，与创建数据集相同，详细参数可参见创建数据集时表格类型的参数说明。

图 2-49 导入数据集-OBS 目录

导入成功后，数据将自动同步到数据集中。您可以在“数据集”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。其中，“表格”类型数据集不支持从Manifest文件导入。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集所在行，单击操作列的“更多 > 导入”。

   或者，您可以单击数据集名称，进入数据集“概览”页，在页面右上角单击“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“添加标签”添加，也可以单击标签右侧的删除图标删除标签。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”：难列指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 2-50 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“数据集”页面，单击数据集的名称进入数据集概览页，然后在概览页右上角单击“开始标注”进去数据集详情页，查看详细数据并进行数据标注。

2.4.2 从 OBS 目录导入的规范说明

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”、“表格”和“声音分类”类型的数据集，支持从OBS目录导入数据。其中，“表格”类型的数据集，支持从OBS、DWS、DLI和MRS等数据源导入数据。

说明

从OBS目录导入数据时，当前操作用户需具备此OBS路径的读取权限。

图像分类

- 图像分类的数据支持两种格式，第一种方式（目录方式）只支持单标签。第二种方式（txt标签文件）支持多标签。
  - 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。当存在多层目录时，则以最后一层目录为标签名。

    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
```

– 当目录下存在对应的txt文件时，以txt文件内容作为图像的标签，优先级高于第一种格式。

  示例如下所示，import-dir-1和imort-dir-2为导入子目录。

```
dataset-import-example
├─import-dir-1
│     10.jpg
│     10.txt
│     11.jpg
│     11.txt
│     12.jpg
│     12.txt
└─import-dir-2
      1.jpg
      1.txt
      2.jpg
      2.txt
```

  单标签的标签文件示例，如1.txt文件内容如下所示：

```
Cat
```

  多标签的标签文件示例，如1.txt文件内容如下所示：

```
Cat
Dog
```

- 只支持JPG、JPEG、PNG、BMP格式的图片。单张图片大小不能超过5MB，且单次上传的图片总大小不能超过8MB。

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2-21**。

  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
```

  标注文件的示例如下所示：

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<annotation>
   <folder>NA</folder>
   <filename>bike_1_1593531469339.png</filename>
   <source>
      <database>Unknown</database>
   </source>
   <size>
      <width>554</width>
      <height>606</height>
      <depth>3</depth>
   </size>
   <segmented>0</segmented>
   <object>
      <name>Dog</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
      <occluded>0</occluded>
      <bndbox>
         <xmin>279</xmin>
```

```
            <ymin>52</ymin>
            <xmax>474</xmax>
            <ymax>278</ymax>
        </bndbox>
    </object>
    <object>
        <name>Cat</name>
        <pose>Unspecified</pose>
        <truncated>0</truncated>
        <difficult>0</difficult>
        <occluded>0</occluded>
        <bndbox>
            <xmin>279</xmin>
            <ymin>198</ymin>
            <xmax>456</xmax>
            <ymax>421</ymax>
        </bndbox>
    </object>
</annotation>
```

- 只支持JPG、JPEG、PNG、BMP格式的图片，单张图片大小不能超过5MB，且单次上传的图片总大小不能超过8MB。

图像分割

- 图像分割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_114746.jpg”，那么标注文件的文件名应为“IMG_20180919_114746.xml”。

  图像分割的标注文件基于PASCAL VOC格式增加了字段mask_source和mask_color，格式详细说明请参见**表2-17**。

  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
```

  标注文件的示例如下所示：

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<annotation>
    <folder>NA</folder>
    <filename>image_0006.jpg</filename>
    <source>
        <database>Unknown</database>
    </source>
    <size>
        <width>230</width>
        <height>300</height>
        <depth>3</depth>
    </size>
    <segmented>1</segmented>
    <mask_source>obs://xianao/out/dataset-8153-Jmf5ylLjRmSacj9KevS/annotation/V001/
segmentationClassRaw/image_0006.png</mask_source>
    <object>
        <name>bike</name>
        <pose>Unspecified</pose>
        <truncated>0</truncated>
        <difficult>0</difficult>
        <mask_color>193,243,53</mask_color>
        <occluded>0</occluded>
        <polygon>
            <x1>71</x1>
            <y1>48</y1>
            <x2>75</x2>
```

```
            <y2>73</y2>
            <x3>49</x3>
            <y3>69</y3>
            <x4>68</x4>
            <y4>92</y4>
            <x5>90</x5>
            <y5>101</y5>
            <x6>45</x6>
            <y6>110</y6>
            <x7>71</x7>
            <y7>48</y7>
        </polygon>
    </object>
</annotation>
```

文本分类

文本分类支持导入“txt”和“csv”两种文件类型，文本的编码格式支持“UTF-8”和“GBK”。

文本分类的标注对象和标注文件有2种存放模式。

- 文本和标注合并：文本分类的标注对象和标注内容在一个文本文件内，标注对象与标注内容之间，多个标注内容之间可分别指定分隔符。

  例如，文本文件的内容如下所示。标注对象与标注内容之间采用tab键分隔。

  ```
  手感很好，反应速度很快，不知道以后怎样	positive
  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	positive
  没充一会电源怎么也会发热呢音量健不好用回弹不好	negative
  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	positive
  ```

- 文本和标注分离：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
  negative
  negative
  positive
  ```

  此数据格式要求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
```

表格

表格支持从4种数据源导入数据，分别为对象存储服务（OBS）、数据仓库服务（DWS）、数据湖探索服务（DLI）、MapReduce服务（MRS）。

导入说明：

1. 导入成功的前提是，数据源的schema需要与创建数据集指定的schema保持一致。其中schema指表格的列名和类型，创建数据集时一旦指定，不支持修改。
2. 数据格式不合法，会将数据置为null，详见**表2-6**。
3. 从OBS或者MRS导入csv文件，不会校验数据类型，但是列数需要跟数据集的schema保持一致。

下面分别介绍如下几种数据源导入：

- 从OBS导入数据

  支持从OBS导入csv文件，需要选择文件所在目录，其中csv文件的列数需要跟数据集schema一致。支持自动获取csv文件的schema。

  ```
  ├─dataset-import-example
  │    table_import_1.csv
  │    table_import_2.csv
  │    table_import_3.csv
  │    table_import_4.csv
  ```

- 从DWS导入数据

  从DWS导入数据，用户需要选择对应的DWS集群，并输入需要对应的数据库名、表名以及用户名和密码。所导入表的schema(列名和类型)需要跟数据集相同。

- 从DLI导入数据

  从DLI导入数据，用户需要选择DLI队列、数据库和表名称。所选择的表的schema(列名和类型)需与数据集一致，支持自动获取所选择表的schema。DLI的default队列只用作体验，不同帐号间可能会出现抢占的情况，需进行资源排队，不能保证每次都可以得到资源执行相关操作。DLI支持schema映射的功能，即导入的表的schema的字段名称可以不和数据集相同，但类型要保持一致。

- 从MRS导入数据

  只支持从分析集群导入数据，流式集群不支持导入。从MRS服务中导入存储在HDFS上的csv格式的数据，首先需要选择已有的MRS集群，并从HDFS文件列表选择文件名称或所在目录，导入文件的列数需与数据集schema一致。

2.4.3 导入 Manifest 文件的规范说明

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。

说明

Manifest文件编写规范要求较多，推荐使用OBS目录导入方式导入新数据。一般此功能常用于不同区域或不同帐号下ModelArts的数据迁移，即当您已在某一区域使用ModelArts完成数据标注，发布后的数据集可从输出路径下获得其对应的Manifest文件。在获取此Manifest文件后，可将此数据集导入其他区域或者其他帐号的ModelArts中，导入后的数据已携带标注信息，无需重复标注，提升开发效率。

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": … }
```

  为了说明方便，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-v20190123113030412 3.manifest”。

图像分类

```
{
    "source":"s3://path/to/image1.jpg",
    "usage":"TRAIN",
    "hard":"true",
    "hard-coefficient":0.8,
    "id":"0162005993f8065ef47eefb59d1e4970",
    "annotation": [
        {
            "type": "modelarts/image_classification",
            "name": "cat",
            "property": {
                "color":"white",
                "kind":"Persian cat"
            },
            "hard":"true",
            "hard-coefficient":0.8,
            "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 2-14 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2-15。
usage	否	默认为空，取值范围： ● 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。
id	否	此参数为系统导出的样本id，导入时可以不用填写。
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2-16。
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 2-15 数据来源类型

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

表 2-16 annotation 对象说明

字段	是否必选	说明
type	是	标签类型。取值范围为： ● 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。

字段	是否必选	说明
id	是/否	标签ID。对于三元组是必选字段，对于其他类型为可选字段。三元组的实体标签ID格式为“E+数字”，比如“E1”、“E2”，三元组的关系标签ID格式为“R+数字”，例如“R1”、“R2”。
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。 ● human
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。

图像分割

```
{
   "annotation": [{
      "annotation-format": "PASCAL VOC",
      "type": "modelarts/image_segmentation",
      "annotation-loc": "s3://path/to/annotation/image1.xml",
      "creation-time": "2020-12-16 21:36:27",
      "annotated-by": "human"
   }],
   "usage": "train",
   "source": "s3://path/to/image1.jpg",
   "id": "16d196c19bf61994d7deccafa435398c",
   "sample-type": 0
}
```

- “source”、“usage”、“annotation”等参数说明与图像分类一致，详细说明请参见表2-14。
- “annotation-loc”：对于图像分割、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”。
- “**sample-type**”：样本格式，0表示图片，1表示文本，2表示语音，4表示表格，6表示视频

表 2-17 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。

字段	是否必选	说明
size	是	表示图像的像素信息。 ● 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。
mask_source	否	表示图像分割保存的mask路径
object	是	表示物体检测信息，多个物体标注会有多个object体。 ● 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2-18**。 ● mask_color：必选字段，标签的颜色，以RGB值表示

表 2-18 标注框类型描述

type	形状	标注信息
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> <x7>100<x7> <y7>100<y7>

示例：

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<annotation>
    <folder>NA</folder>
    <filename>image_0006.jpg</filename>
    <source>
        <database>Unknown</database>
    </source>
    <size>
        <width>230</width>
        <height>300</height>
        <depth>3</depth>
    </size>
    <segmented>1</segmented>
    <mask_source>obs://xianao/out/dataset-8153-Jmf5ylLjRmSacj9KevS/annotation/V001/
segmentationClassRaw/image_0006.png</mask_source>
    <object>
        <name>bike</name>
        <pose>Unspecified</pose>
        <truncated>0</truncated>
        <difficult>0</difficult>
        <mask_color>193,243,53</mask_color>
        <occluded>0</occluded>
        <polygon>
            <x1>71</x1>
            <y1>48</y1>
            <x2>75</x2>
            <y2>73</y2>
            <x3>49</x3>
            <y3>69</y3>
            <x4>68</x4>
            <y4>92</y4>
            <x5>90</x5>
            <y5>101</y5>
            <x6>45</x6>
            <y6>110</y6>
```

```
            <x7>71</x7>
            <y7>48</y7>
        </polygon>
    </object>
</annotation>
```

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"
        } ]
}
```

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**图像分类**相同，请参见**表2-14**。

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**图像分类**一致，详细说明请参见**表2-14**。

其中，property的参数解释如**表2-19**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 2-19 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。

文本三元组

```
{
    "source":"content://"Three Body" is a series of long science fiction novels created by Liu Cix.",
    "usage":"TRAIN",
    "annotation":[
        {
            "type":"modelarts/text_entity",
            "name":"Person",
            "id":"E1",
            "property":{
                "@modelarts:start_index":67,
                "@modelarts:end_index":74
            },
            "annotated-by":"human",
            "creation-time":"2019-01-23 11:30:30"
        },
        {
            "type":"modelarts/text_entity",
            "name":"Book",
            "id":"E2",
            "property":{
                "@modelarts:start_index":0,
                "@modelarts:end_index":12
            },
            "annotated-by":"human",
            "creation-time":"2019-01-23 11:30:30"
        },
        {
            "type":"modelarts/text_triplet",
            "name":"Author",
            "id":"R1",
            "property":{
                "@modelarts:from":"E1",
                "@modelarts:to":"E2"
            },
            "annotated-by":"human",
            "creation-time":"2019-01-23 11:30:30"
        },
        {
            "type":"modelarts/text_triplet",
            "name":"Works",
            "id":"R2",
            "property":{
                "@modelarts:from":"E2",
                "@modelarts:to":"E1"
            },
            "annotated-by":"human",
            "creation-time":"2019-01-23 11:30:30"
        }
    ]
}
```

“source”、“usage”、“annotation”等参数说明与**图像分类**一致，详细说明请参见**表2-14**。

其中，property的参数解释如表5 property参数说明所示。其中，“@modelarts:start_index”和“@modelarts:end_index”和文本命名实体的参数说明一致。例如，当“source”："content://"Three Body" is a series of long science fiction novels created by Liu Cix."时，“Liu Cix”是实体Person（人物），“Three Body”是实体Book（书籍），Person指向Book的关系是Author（作者），Book指向Person的关系是Works（作品）。

表 2-20 property 参数说明

参数名	数据类型	说明
@modelarts:start_index	Integer	三元组实体的起始位置，值从0开始，包括start_index所指的字符。
@modelarts:end_index	Integer	三元组实体的结束位置，但不包括end_index所指的字符。
@modelarts:from	String	三元组关系的起始实体id
@modelarts:to	String	三元组关系的指向实体id

物体检测

```
{
    "source":"s3://path/to/image1.jpg",
    "usage":"TRAIN",
    "hard":"true",
    "hard-coefficient":0.8,
    "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”等参数说明与图像分类一致，详细说明请参见表2-14。
- “annotation-loc”：对于物体检测、图像分割是必选字段，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 2-21 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。

字段	是否必选	说明
size	是	表示图像的像素信息。 ● 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。 ● 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2-22。

表 2-22 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>
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>

type	形状	标注信息
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>
    <hard-coefficient>0.8</hard-coefficient>
    <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”等参数说明与图像分类一致，详细说明请参见表2-14。

语音内容

```
{
   "source":"s3://path/to/audio1.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”等参数说明与图像分类一致，详细说明请参见表2-14。
- “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/audio1.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”等参数说明与图像分类一致，详细说明请参见表2-14。
- “property”的参数解释如表2-23所示。

表 2-23 “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。

参数名	数据类型	描述
@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。

视频标注

```
{
    "annotation": [{
        "annotation-format": "PASCAL VOC",
        "type": "modelarts/object_detection",
        "annotation-loc": "s3://path/to/annotation1_t1.473722.xml",
        "creation-time": "2020-10-09 14:08:24",
        "annotated-by": "human"
    }],
    "usage": "train",
    "property": {
        "@modelarts:parent_duration": 8,
        "@modelarts:parent_source": "s3://path/to/annotation1.mp4",
        "@modelarts:time_in_video": 1.473722
    },
    "source": "s3://input/path/to/annotation1_t1.473722.jpg",
    "id": "43d88677c1e9a971eeb692a80534b5d5",
    "sample-type": 0
}
```

- “source”、“usage”、“annotation”等参数说明与图像分类一致，详细说明请参见表2-14。
- “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”。
- “**sample-type**”：样本格式，0表示图片，1表示文本，2表示语音，4表示表格，6表示视频。

表 2-24 property 参数说明

参数名	数据类型	说明
@modelarts:parent_duration	Double	标注视频的时长，单位：秒。
@modelarts:time_in_video	Double	标注的视频帧的时间戳，单位：秒。
@modelarts:parent_source	String	标注视频的OBS路径。

表 2-25 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。 ● 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。 ● 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2-26**。

表 2-26 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>
line	线	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>200<y2>

type	形状	标注信息
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_t1.473722.jpg.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>
    <hard-coefficient>0.8</hard-coefficient>
    <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>
```

2.5 导出数据

针对数据集中的数据，包含“已标注”和“未标注”的数据。您可以选中需要部分图片或者通过筛选条件筛选出需要的数据，导出成新的数据集，或者将数据导出至指定的OBS目录下，您可以通过任务历史查看数据导出的历史记录。

说明

目前只有“图像分类”、“物体检测”、“图像分割”、“自由格式”类型的数据集支持导出功能。

- “图像分类”只支持导出txt格式的标注文件。
- “物体检测”只支持导出Pascal VOC格式的XML标注文件。
- “图像分割”只支持导出Pascal VOC格式的XML标注文件以及Mask图像。
- “自由格式”直接导出数据集的所有样本文件。

导出新数据集

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，选择“物体检测”或“图像分类”类型的数据集，单击数据集名称进入“数据集概览页”。

   说明

   针对“自由格式”类型的数据集，单击数据集名称即可进入数据集详情页，直接跳转至步骤4。

3. 在“数据集概览页”，单击右上角“开始标注”，进入数据集详情页。
4. 在数据集详情页面中，选中导出数据或者筛选出数据，然后单击“导出 > 新数据集”。

   图 2-51 选择图片或筛选图片

5. 在弹出的“导出新数据集”对话框中，填写相关信息，然后单击“确定”，开始执行导出操作。

   “名称”：新数据集名称。

   “保存路径”：表示新数据集的输入路径，即当前数据导出后存储的OBS路径。

   “输出路径”：表示新数据集的输出路径，即新数据集在完成标注后输出的路径。“输出路径”不能与“保存路径”为同一路径，且“输出路径”不能是“保存路径”的子目录。

   “导出范围”：“导出当前选中样本”，或者“导出当前筛选条件下的所有样本”。

   “开启难例属性”：设置是否开启难例。

图 2-52 导出新数据集

6. 数据集导出成功后，您可以前往“数据集”列表中，查看到新的数据集。

导出数据至 OBS

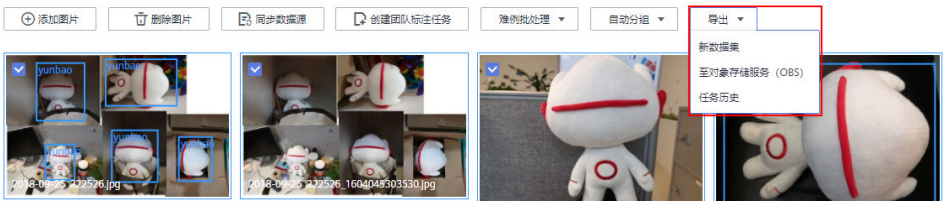
1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，选择“物体检测”或“图像分类”类型的数据集，单击数据集名称进入“数据集概览页”。

   说明

   针对“自由格式”类型的数据集，单击数据集名称即可进入数据集详情页，直接跳转至步骤4。

3. 在“数据集概览页”，单击右上角“开始标注”，进入数据集详情页。
4. 在数据集详情页面中，选中导出数据或者筛选出数据，然后单击“导出 > 至对象存储服务（OBS）”。

图 2-53 选择图片或筛选图片导出



5. 在弹出的“导出至对象存储服务（OBS）”对话框中，填写相关信息，然后单击“确定”，开始执行导出操作。

   “保存路径”：即导出数据存储的路径。建议不要将数据存储至当前数据集所在的输入路径或输出路径。

   “导出范围”：“导出当前选中样本”，或者“导出当前筛选条件下的所有样本”。

   “开启难例属性”：设置是否开启难例。

图 2-54 导出至 OBS

6. 数据导出成功后，您可以前往您设置的保存路径，查看到存储的数据。

任务历史

当您导出新数据集或导出数据至OBS，您可以通过任务历史查看导出任务明细。

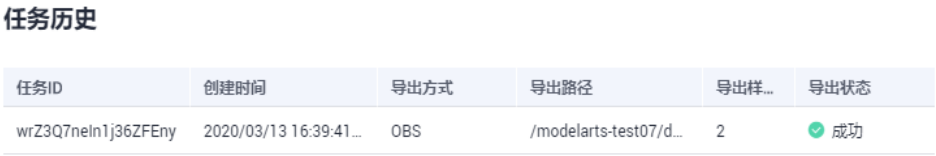
1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，选择“物体检测”或“图像分类”类型的数据集，单击数据集名称进入“数据集概览页”。

   说明

   针对“自由格式”类型的数据集，单击数据集名称即可进入数据集详情页，直接跳转至步骤4。

3. 在“数据集概览页”，单击右上角“开始标注”，进入数据集详情页。
4. 在数据集详情页面中，选中导出数据或者筛选出数据，然后单击“导出 > 任务历史”。
5. 在弹出的“任务历史”对话框中，可以查看该数据集之前的导出任务历史。包括“任务ID”、“创建时间”、“导出方式”、“导出路径”、“导出样本总数”和“导出状态”。

图 2-55 任务历史



2.6 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“数据管理>数据集”，进入“数据集”管理页面。

2. 在数据集列表中，单击操作列的“更多 > 修改”。
   或者，您可以单击数据集名称，进入数据集“概览”页，在页面右上角单击“修改”。
3. 参考表2-27修改数据集基本信息，然后单击“确定”完成修改。

图 2-56 修改数据集

表 2-27 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、下划线或者中划线组成的合法字符串。
描述	数据集的简要描述。
标签集	针对不同类型数据集，标签集的设置不同，请参考创建数据集（旧版）中不同类型数据集的详细参数指导进行修改。当前对标签集个数没有限制。

2.7 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递增规则进行命名，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。
- 针对每一个数据集版本，您可以通过“存储路径”参数，获得此版本对应的Manifest文件格式的数据集。可用于导入数据或难例筛选操作。
- 表格数据集暂不支持切换版本。

发布数据集

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，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”。

   或者，您可以单击数据集名称，进入数据集“概览”页，在页面右上角单击“发布”。
3. 在“发布新版本”弹出框中，填写发布数据集的相关参数，然后单击“确定”。

表 2-28 发布数据集的参数说明

参数	描述
“版本名称”	默认按V001、V002递增规则进行命名，您也可以自定义版本名称。版本名称只能包含字母、数字、中划线或下划线。
“版本格式”	仅“表格”类型数据集支持设置版本格式，支持“CSV”和“CarbonData”两种。 **说明** 如果导出的CSV文件中存在以“=”“+”“-”和“@”开头的命令时，为了安全考虑，ModelArts会自动加上Tab键，并对双引号进行转义处理。
“数据切分”	仅“图像分类”、“物体检测”、“文本分类”和“声音分类”类型数据集支持进行数据切分功能。 默认不启用。启用后，需设置对应的训练验证比例。 输入“训练集比例”，数值只能是0~1区间内的数。设置好“训练集比例”后，“验证集比例”自动填充。“训练集比例”加“验证集比例”等于1。 “训练集比例”即用于训练模型的样本数据比例；“验证集比例”即用于验证模型的样本数据比例。“训练验证比例”会影响训练模板的性能。
“描述”	针对当前发布的数据集版本的描述信息。
“开启难例属性”	仅“图像分类”和“物体检测”类型数据集支持难例属性。 默认不开启。启用后，会将此数据集的难例属性等信息写入对应的Manifest文件中。

图 2-57 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
         |-- VersionMame1
            |-- VersionMame1.manifest
         |-- VersionMame2
            ...
         |-- ...
```

以物体检测为例，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
            |-- annotation
               |-- file1.xml
         |-- VersionMame2
            ...
         |-- ...
```

以视频标注为例，在数据集发布后，标注结果将标注结果文件（XML）存放在数据集输出目录下。

```
|-- user-specified-output-path
    |-- DatasetName-datasetId
       |-- annotation
          |-- VersionMame1
```

```
            |-- VersionMame1.manifest
            |-- annotations
              |-- images
                |-- videoName1
                  |-- videoName1.timestamp.xml
                |-- videoName2
                  |-- videoName2.timestamp.xml
        |-- VersionMame2
            ...
        |-- ...
```

视频标注的关键帧存在数据集的输入目录下。

```
|-- user-specified-input-path
    |-- images
      |-- videoName1
          |-- videoName1.timestamp.jpg
      |-- videoName2
          |-- videoName2.timestamp.jpg
```

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“数据管理>数据集”页面中，单击数据集操作列的“更多 > 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。

2.9 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“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，单击操作列的“更多 > 版本管理”，进入数据集“版本管理”页面。

   您可以查看数据集的基本信息，并在左侧查看版本及其发布时间。

图 2-58 查看数据集版本

设置当前版本

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，单击操作列的“更多 > 版本管理”，进入数据集“版本管理”页面。
3. 在“版本管理”页面中，选择对应的数据集版本，在数据集版本基本信息区域，单击“设置为当前版本”。设置完成后，版本名称右侧将显示为“当前版本”。

说明

只有状态为“正常”的版本，才能被设置为当前版本。

图 2-59 设置当前版本

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“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，单击操作列的“更多 > 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。

2.10 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。

背景信息

- 目前只有“图像分类”和“物体检测”类型的数据集支持智能标注功能。
- 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。
- 检查用于标注的图片数据，确保您的图片数据中，不存在RGBA四通道图片。如果存在四通道图片，智能标注任务将运行失败，因此，请从数据集中删除四通道图片后，再启动智能标注。

智能标注

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理 > 数据集”，进入“数据集”管理页面。
2. 在数据集列表中，选择“物体检测”或“图像分类”类型的数据集，单击操作列的“智能标注”启动智能标注作业。
3. 在弹出的“启动智能标注”对话框中，选择智能标注类型，可选“主动学习”或者“预标注”，详见表2-29和表2-30。

表 2-29 主动学习

参数	说明
智能标注类型	“主动学习”。“主动学习”表示系统将自动使用半监督学习、难例筛选等多种手段进行智能标注，降低人工标注量，帮助用户找到难例。
算法类型	针对“图像分类”类型的数据集，您需要选择以下参数。 “快速型”：仅使用已标注的样本进行训练。 “精准型”：会额外使用未标注的样本做半监督训练，使得模型精度更高。

表 2-30 预标注

参数	说明
智能标注类型	“预标注”。“预标注”表示选择用户AI应用管理里面的AI应用,选择模型时需要注意模型类型和数据集的标注类型相匹配。预标注结束后，如果标注结果符合平台定义的标准标注格式，系统将进行难例筛选，该步骤不影响预标注结果。
选择模型及版本	• “我的AI应用”。您可以根据实际需求选择您的AI应用。您需要在目标AI应用的左侧单击下拉三角标，选择合适的版本。您的AI应用导入参见**创建AI应用**。 • “我的订阅”。您可以根据实际需求选择AI Gallery中已订阅的AI应用。您需要在目标AI应用的左侧单击下拉三角标，选择合适的版本。查找AI应用参见**从Gallery订阅模型**。
计算节点规格	在下拉框中，您可以选择目前ModelArts支持的节点规格选项。
计算节点个数	默认为1。您可以根据您的实际情况选择，最大为5。

说明

- 针对“物体检测”类型的数据集，选择“主动学习”时，只支持识别和标注矩形框。
- 当系统中智能标注作业过多时，可能会出现排队的情况，导致作业一直处于“标注中”的状态。请您耐心等待，系统会按照顺序完成标注作业。

图 2-60 启动智能标注（图像分类）

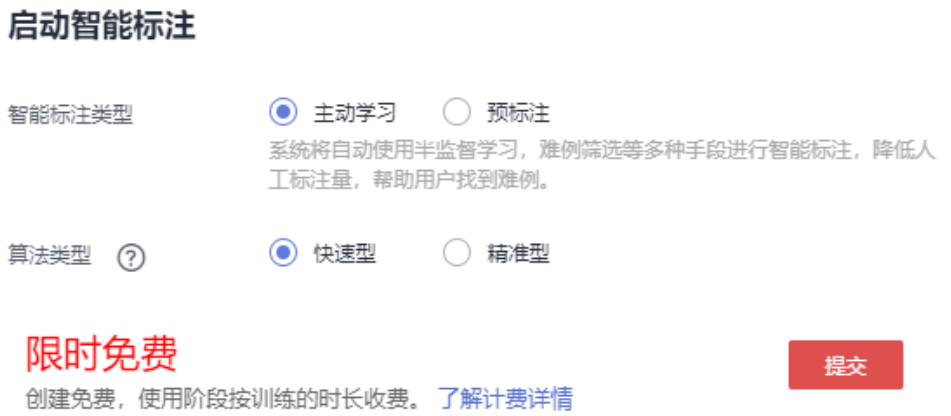


图 2-61 启动智能标注（物体检测）

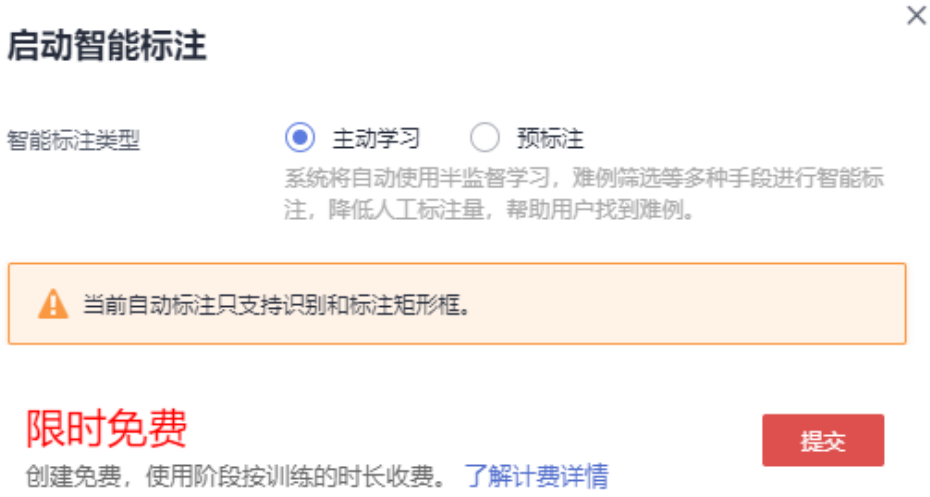


图 2-62 启动智能标注（预标注）

4. 完成参数设置后，单击“提交”，即可启动智能标注。
5. 在数据集列表中，单击数据集名称进入“数据集概览”页。
6. 在“数据集概览页”，单击右上角“开始标注”，进入数据集详情页。
7. 在数据集详情页，单击“待确认”页签，即可查看智能标注进度。
   您也可以在该页签，“启动智能标注”或者查看“智能标注历史”。

图 2-63 标注进度

8. 智能标注完成后，“待确认”页面将呈现所有标注后的图片列表。
   - 图像分类数据集

     在“待确认”页面查看标签是否准确，勾选标注准确的图片，然后单击“确认”完成智能标注结果的确认。确认完成后的图片将被归类至“已标注”页面下。

     针对标为“难例”的图片，您可以根据实际情况判断，手工修正标签。详细操作及示例请参见**•针对“图像分类”数据集**。
   - 物体检测数据集

     在“待确认”页面，单击图片查看标注详情，查看标签及目标框是否准确，针对标注准确的图片单击“确认标注”完成智能标注结果的确认。确认完成后的图片将被归类至“已标注”页面下。

     针对标为“难例”的图片，您可以根据实际情况判断，手工修正标签或目标框。详细操作及示例请参见**•针对“物体检测”数据集**。

2.11 难例确认

在数据量很大的标注任务中，标注初期由于已标注图片不足，智能标注的结果无法直接用于训练。若对所有的未标注数据一一进行调整确认仍然需要较大的人力和时间成本。为了更快地完成标注任务，在对未标注数据进行智能标注的任务中，ModelArts嵌入了自动难例发现功能。该功能会对剩余未标注图片的标注优先级给出建议。因为标注优先级高的图片的智能标注结果未达到预期，所以称之为难例。

ModelArts平台提供的自动难例发现功能，在智能标注以及数据采集筛选过程中，将自动标注出难例，建议对难例数据进一步确认标注，然后将其加入训练数据集中，使用此数据集训练模型，可得到精度更高的模型。首先，针对智能标注和采集筛选任务，难例的发现操作是系统自动执行的，无需人工介入，仅需针对标注后的数据进行确认和修改即可，提升数据管理和标注效率。其次，您可以基于难例的情况，补充类似数据，提升数据集的丰富性，进一步提升模型训练的精度。在数据集管理中，对难例的管理有如下场景。

- **智能标注后，确认难例**
- **将数据集中的数据标注为难例**

说明

目前只有“图像分类”和“物体检测”类型的数据集支持难例发现功能。

智能标注后，确认难例

“智能标注”任务执行过程中，ModelArts将自动识别难例，并完成标注。当智能标注结束后，难例标注结果将呈现在“待确认”页签，建议您对难例数据进行人工修正，然后确认标注。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，选择“物体检测”或“图像分类”类型的数据集，单击数据集名称进入“数据集概览页”。
3. 在“数据集概览页”，单击右上角“开始标注”，进入数据集详情页。
4. 在数据集详情页，单击“待确认”页签，查看并确认难例。

说明

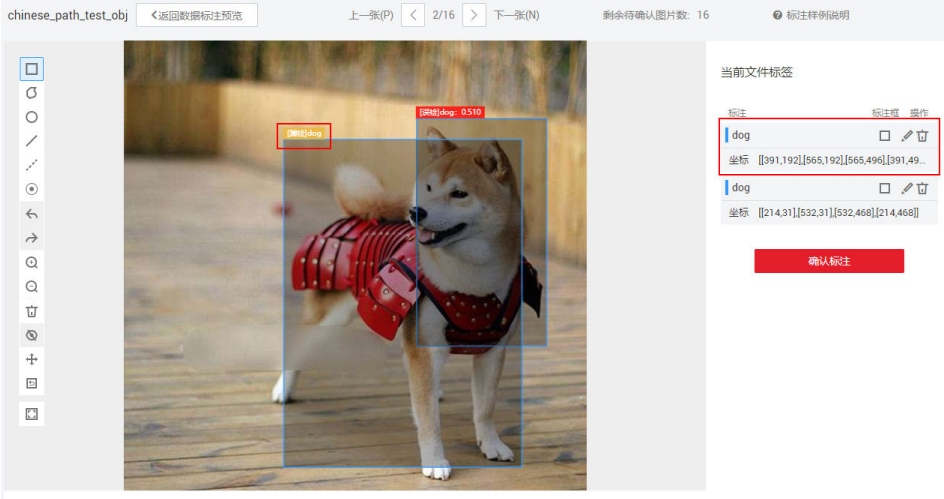
只有当智能标注任务完成后，待确认页签才会显示标注数据。否则，此页签内容为空。智能标注操作请参见**智能标注**。

– 针对“物体检测”数据集

在“待确认”页签中，单击图片展开标注详情，查看图片数据的标注情况，如标签是否准确、目标框位置添加是否准确。如果智能标注结果不准确，建议手工调整标签或目标框，然后单击“确认标注”。完成确认后，重新标注的数据将呈现在“已标注”页签下。

如**图2-64**所示的难例，dog标签的目标框位置不准确，使用标注框重新标注，如图中的“漏检”目标框，然后需要将原先标注错误的目标框删除，即“误检”标签框。手工调整后，单击“确认标注”完成难例确认。

图 2-64 物体检测的难例确认

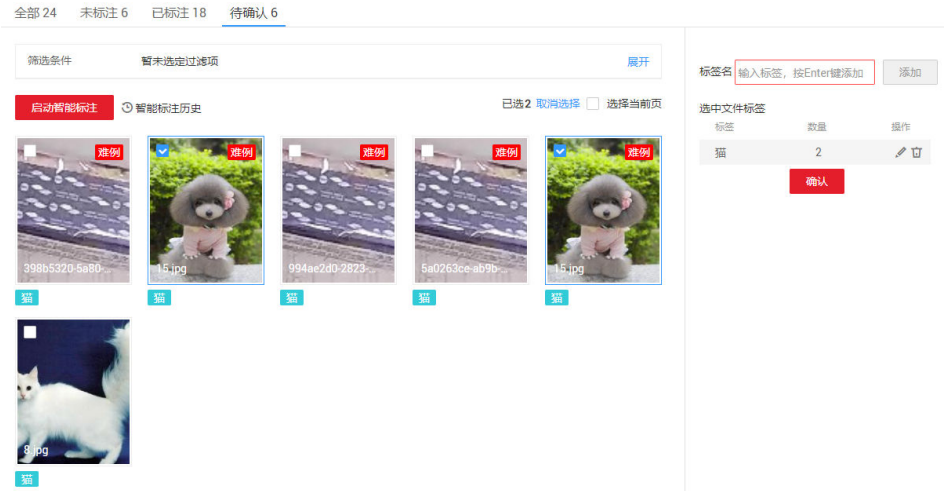


– 针对“图像分类”数据集

在“待确认”页签中，查看标注难例的图片，其添加的标签是否准确。勾选标注不准确的图片，删除错误标签，然后在右侧“标签名”处添加准确标签。单击“确认”，勾选的图片及其标注情况，将呈现在“已标注”页签下。

如**图2-65**所示，选中的图片为标注错误图片，在右侧删除错误标签，然后在标签名处添加“狗”的标签，然后单击“确认”，完成难例确认。

图 2-65 图像分类的难例确认

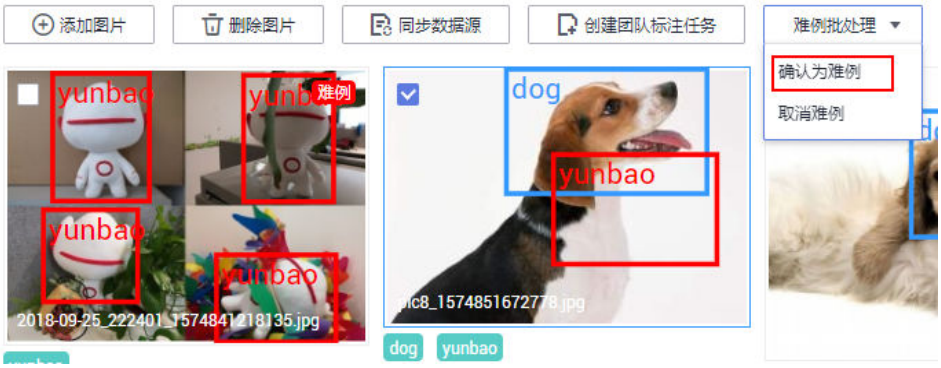


将数据集中的数据标注为难例

针对数据集中，已标注或未标注数据，也可以将图片数据标注为难例。标注为难例的数据，对后续模型训练中，通过内置规则提升模型精度。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，选择“物体检测”或“图像分类”类型的数据集，单击数据集名称进入“数据集概览页”。
3. 在“数据集概览页”，单击右上角“开始标注”，进入数据集详情页。
4. 在数据集详情页，单击“已标注”、“未标注”或“全部”页签，勾选需标注为难例的图片，然后单击“难例批处理 > 确认为难例”。完成标注后，图片预览时，其右上角将显示为“难例”。

图 2-66 确认为难例



2.12 自动分组

为了提升智能标注算法精度，可以均衡标注多个类别，有助于提升智能标注算法精度。ModelArts内置了分组算法，您可以针对您选中的数据，执行自动分组，提升您的数据标注效率。

自动分组可以理解为数据标注的预处理，先使用聚类算法对未标注图片进行聚类，再根据聚类结果进行处理，可以分组打标或者清洗图片。

例如，用户通过搜索引擎搜索XX，将相关图片下载并上传到数据集，然后再使用自动分组，可以将XX图片分类，比如论文、宣传海报、确认为XX的图片、其他。用户可以根据分组结果，快速剔除掉不想要的，或者将某一类直接全选后添加标签。

说明

目前只有“图像分类”、“物体检测”和“图像分割”类型的数据集支持自动分组功能。

启动自动分组任务

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 在数据集列表中，选择“物体检测”或“图像分类”类型的数据集，单击数据集名称进入“数据集概览页”。
3. 在“数据集概览页”，单击右上角“开始标注”，进入数据集详情页。
4. 在数据集详情页的“全部”页签中，单击“自动分组 > 启动任务”。

   说明

   只能在“全部”页签下启动自动分组任务或查看任务历史。

5. 在弹出的“自动分组”对话框中，填写参数信息，然后单击“确定”。
   - “分组数”：填写2~200之间的整数，指将图片分为多少组。
   - “结果处理方式”：“更新属性到当前样本中”，或者“保存到对象存储服务（OBS）”。
   - “属性名称”：当选择“更新属性到当前样本中”时，需输入一个属性名称。
   - “结果存储目录”：当选择“保存到对象存储服务（OBS）”时，需指定一个用于存储的OBS路径。
   - “高级特征选项”：启用此功能后，可选择“清晰度”、“亮度”、“图像色彩”等维度为自动分组功能增加选项，使得分组着重于图片亮度、色彩和清晰度等特征进行分组。支持多选。

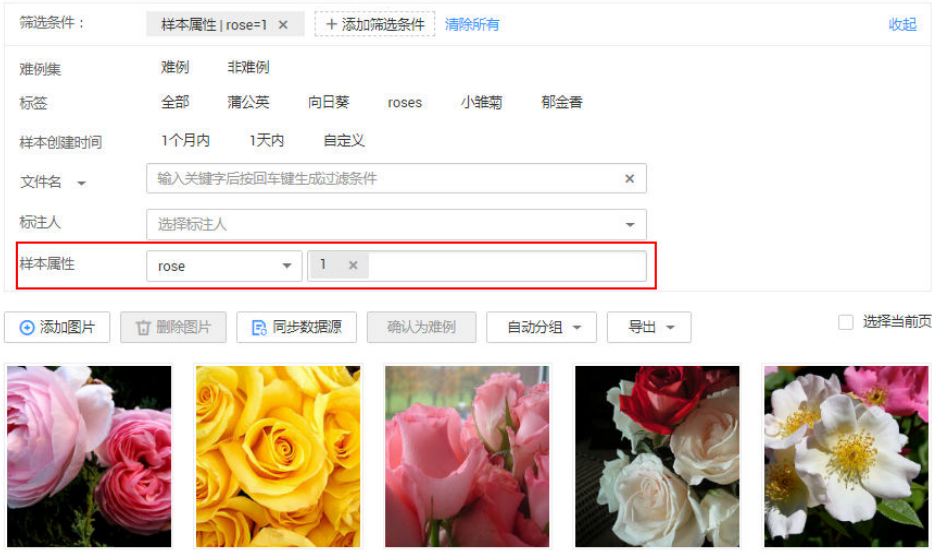
图 2-67 自动分组

6. 启动任务提交成功后，界面右上角显示此任务的进度。等待任务执行完成后，您可以查看自动分组任务的历史记录，了解任务状态。

查看自动分组结果

在数据集详情页面的“全部”页签中，展开“筛选条件”，将“样本属性”设置为自动分组任务中的“属性名称”，并通过设置样本属性值，筛选出分组结果。

图 2-68 查看自动分组结果



查看自动分组的历史任务

在数据集详情页面的“全部”页签中，单击“自动分组 > 任务历史”。在弹出的“任务历史”对话框中，展示当前数据集之前执行的自动分组任务的基本信息。

图 2-69 自动分组任务历史

2.13 数据特征

基于图片或目标框对图片的各项特征，如模糊度、亮度进行分析，并绘制可视化曲线，帮助处理数据集。

您还可以选择数据集的多个版本，查看其可视化曲线，进行对比分析。

背景信息

- 只有“物体检测”和“图像分类”的数据集支持数据特征分析。
- 只有发布后的数据集支持数据特征分析。发布后的Default格式数据集版本支持数据特征分析。
- 数据特征分析的数据范围，不同类型的数据集，选取范围不同：
  - 在“物体检测”的数据集中，当已标注样本数为0时，发布版本后，数据特征页签版本置灰不可选，无法显示数据特征。有标注后，发布版本，显示已标注的图片的数据特征。
  - 在“图像分类”的数据集中，当已标注样本数为0时，发布版本后，数据特征页签版本置灰不可选，无法显示数据特征。有标注后，发布版本，显示全部的图片的数据特征。
- 数据集中的图片数量要达到一定量级才会具有意义，一般来说，需要有大约1000+的图片。
- “图像分类”支持分析指标有：“分辨率”、“图片高宽比”、“图片亮度”、“图片饱和度”、“清晰度”和“图像色彩的丰富程度”。“物体检测”支持所有的分析指标。目前ModelArts支持的所有分析指标请参见**支持分析指标及其说明**。

数据特征分析

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理>数据集”，进入“数据集”管理页面。
2. 选择对应的数据集，单击操作列的“数据特征”，进入数据集概览页的数据特征页面。

   您也可以在单击数据集名称进入数据集概览页后，单击“数据特征”页签进入。
3. 由于发布后的数据集不会默认启动数据特征分析，针对数据集的各个版本，需手动启动特征分析任务。在数据特征页签下，单击“特征分析”。

   图 2-70 选择特征分析

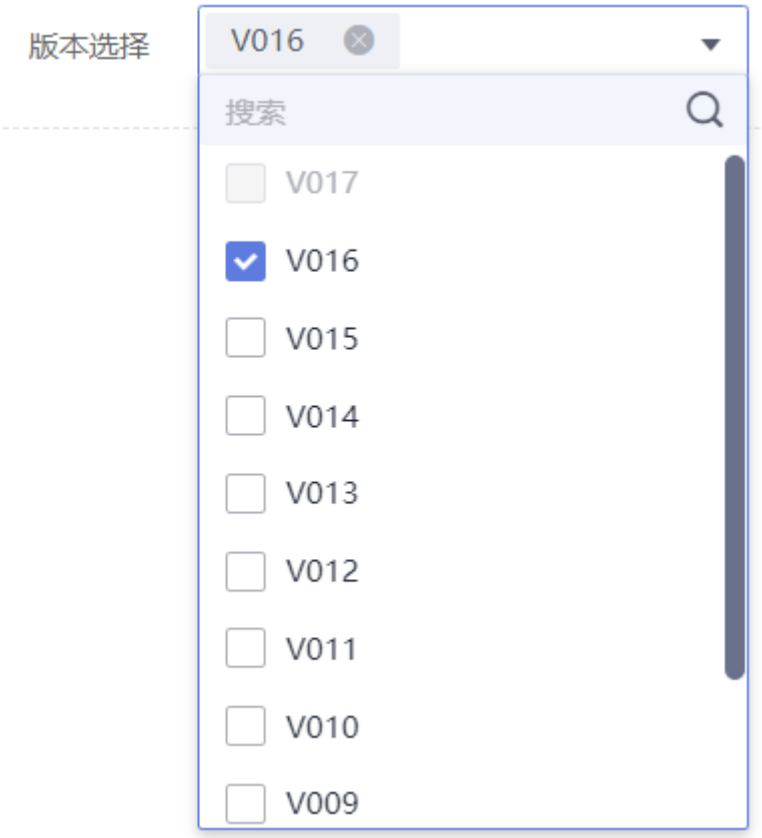
4. 在弹出的对话框中配置需要进行特征分析的数据集版本，然后单击“确定”启动分析。

   “版本选择”，即选择当前数据集的已发布版本。

   图 2-71 启动数据特征分析任务

5. 数据特征分析任务启动后，需执行一段时间，根据数据量不同等待时间不同，请耐心等待。当您选择分析的版本出现在“版本选择”列表下，且可勾选时，即表示分析已完成。

图 2-72 可选择已执行特征分析的版本



6. 查看数据特征分析结果。

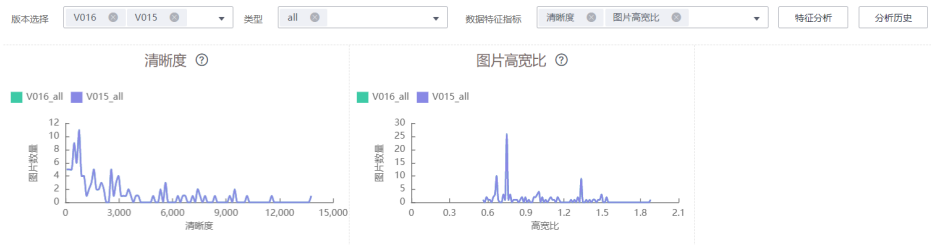
“版本选择”：在右侧下拉框中选择进行对比的版本。也可以只选择一个版本。

“类型”：选择需要分析的类型。支持“all”、“train”、“eval”和“inference”。

“数据特征指标”：在右侧下拉框中勾选需要展示的指标。详细指标说明请参见**支持分析指标及其说明**。

选择完成后，页面将自动呈现您选择对应版本及其指标数据，如**图2-73**所示，您可以根据呈现的图表了解数据分布情况，帮助您更好的处理您的数据。

图 2-73 数据特征分析



7. 查看分析任务的历史记录。

在数据特征分析后，您可以在“数据特征”页签下，单击右侧“任务历史”，可在弹出对话框中查看历史分析任务及其状态。

图 2-74 任务历史

支持分析指标及其说明

表 2-31 分析指标列表

名称	说明	分析说明
分辨率 Resolution	图像分辨率。此处使用面积值作为统计值。	通过指标分析结果查看是否有偏移点。如果存在偏移点，可以对偏移点做resize操作或直接删除。
图片高宽比 Aspect Ratio	图像高宽比，即图片的高度/图片的宽度。	一般呈正态分布，一般用于比较训练集和真实场景数据集的差异。
图片亮度 Brightness	图片亮度，值越大代表观感上亮度越高。	一般呈正态分布，可根据分布中心判断数据集整体偏亮还是偏暗。可根据使用场景调整，比如使用场景是夜晚，图片整体应该偏暗。

名称	说明	分析说明
图片饱和度 Saturation	图片的色彩饱和度，值越大表示图片整体色彩越容易分辨。	一般呈正态分布，一般用于比较训练集和真实场景数据集的差异。
清晰度 Clarity	图片清晰程度，使用拉普拉斯算子计算所得，值越大代表边缘越清晰，图片整体越清晰。	可根据使用场景判断清晰度是否满足需要。比如使用场景的数据采集来自高清摄像头，那么清晰度对应的需要高一些。可通过对数据集做锐化或模糊操作，添加噪声对清晰度做调整。
图像色彩的丰富程度 Colorfulness	横坐标：图像的色彩丰富程度，值越大代表色彩越丰富。 纵坐标：图片数量。	是观感上的色彩丰富程度，一般用于比较训练集和真实场景数据集的差异。
按单张图片中框的个数统计图片分布 Bounding Box Quantity	横坐标：单张图片中框的个数。 纵坐标：图片数量。	对模型而言一张图片的框个数越多越难检测，需要越多的这种数据用作训练。
按单张图片中框的面积标准差统计图片分布 Standard Deviation of Bounding Boxes Per Image	横坐标：单张图片中框的标准差。单张图片只有一个框时，标准差为0。标准差的值越大，表示图片中框大小不一程度越高。 纵坐标：图片数量。	对模型而言一张图中框如果比较多且大小不一，是比较难检测的，可以根据场景添加数据用作训练，或者实际使用没有这种场景可直接删除。
按高宽比统计框数量的分布 Aspect Ratio of Bounding Boxes	横坐标：目标框的高宽比。 纵坐标：框数量（统计所有图片中的框）。	一般呈泊松分布，但与使用场景强相关。多用于比较训练集和验证集的差异，如训练集都是长方形框的情况下，验证集如果是接近正方形的框会有比较大影响。
按面积占比统计框数量的分布 Area Ratio of Bounding Boxes	横坐标：目标框的面积占比，即目标框的面积占整个图片面积的比例，越大表示物体在图片中的占比越大。 纵坐标：框数量（统计所有图片中的框）。	主要判断模型中使用的anchor的分布，如果目标框普遍较大，anchor就可以选择较大。
按边缘化程度统计框数量的分布 Marginalization Value of Bounding Boxes	横坐标：边缘化程度，即目标框中心点距离图片中心点的距离占图片总距离的比值，值越大表示物体越靠近边缘。 纵坐标：框数量（统计所有图片中的框）。	一般呈正态分布。用于判断物体是否处于图片边缘，有一些只露出一部分的边缘物体，可根据需要添加数据集或不标注。

名称	说明	分析说明
按堆叠度统计框数量的分布 Overlap Score of Bounding Boxes	横坐标：堆叠度，单个框被其他的框重叠的部分，取值范围为0～1，值越大表示被其他框覆盖的越多。 纵坐标：框数量（统计所有图片中的框）。	主要用于判断待检测物体的堆叠程度，堆叠物体一般对于检测难度较高，可根据实际使用需要添加数据集或不标注部分物体。
按亮度统计框数量的分布 Brightness of Bounding Boxes	横坐标：目标框的图片亮度，值越大表示越亮。 纵坐标：框数量（统计所有图片中的框）。	一般呈正态分布。主要用于判断待检测物体的亮度。在一些特殊场景中只有物体的部分亮度较暗，可以看是否满足要求。
按清晰度统计框数量的分布 Clarity of Bounding Boxes	横坐标：目标框的清晰度，值越大表示越清晰。 纵坐标：框数量（统计所有图片中的框）。	主要用于判断待检测物体是否存在模糊的情况。比如运动中的物体在采集中可能变得模糊，需要重新采集。

2.14 团队标注

2.14.1 团队标注简介

数据标注任务中，一般由一个人完成，但是针对数据集较大时，需要多人协助完成。ModelArts提供了团队标注功能，可以由多人组成一个标注团队，针对同一个数据集进行标注管理。

说明

团队标注功能当前仅支持“图像分类”、“物体检测”、“文本分类”、“命名实体”、“文本三元组”、“语音分割”类型的数据集。

如何启用团队标注

- 创建数据集时，打开“启用团队标注”开关，同时指定一个标注团队，或者指定标注管理员。

图 2-75 创建数据集时启用

- 对于已创建，且未启用团队标注的数据集。可通过创建一个团队标注任务，直接启用团队标注功能。创建团队标注的操作详情请参见**创建团队标注任务**。

图 2-76 在数据集列表中创建团队标注任务

图 2-77 在数据集概览页中创建团队标注任务

图 2-78 在数据集详情页创建团队标注任务

说明

只有当创建团队标注任务时，标注人员才会收到邮件。创建标注团队及添加标注团队的成员并不会发送邮件。此外，当所有样本都是已标注状态时，创建团队标注任务也不会收到邮件。

团队标注相关操作

- 管理团队
- 管理成员
- 管理团队标注任务

2.14.2 管理团队

团队标注功能是以团队为单位进行管理，数据集启用团队标注功能时，必须指定一个团队。一个团队可以添加多个成员。

背景说明

- 一个帐号最多可添加10个团队。
- 如果数据集需要启用团队标注功能，当前帐号至少拥有一个团队。如果没有，请执行添加团队操作添加。

添加团队

1. 在ModelArts管理控制台左侧导航栏中，选择“数据管理>标注团队”，进入“标注团队”管理页面。
2. 在“标注团队”管理页面，单击“添加团队”。
3. 在弹出的“添加团队”对话框中，填写团队“名称”和“描述”，然后单击“确定”。完成标注团队的添加。

图 2-79 添加团队

团队添加完成后，“标注团队”管理页面呈现新添加的团队，在页面右侧区域，可以查看团队详情。新添加的团队，其成员列表为空，请参考添加成员操作，为您的团队添加成员。

删除团队

当已有的团队不再使用，您可以执行删除操作。

在“标注团队”管理页面中，选中需删除的团队，然后单击“删除”。在弹出的对话框中，确认信息无误后，单击“确定”完成团队删除。

图 2-80 删除团队

2.14.3 管理成员

新添加的团队，其成员列表为空。您需要根据实际情况添加即将参与标注任务的成员信息。

一个团队最多支持添加100个成员，当超过100时，建议分为多个团队进行管理。

添加成员

1. 在ModelArts管理控制台左侧导航栏中，选择“数据管理>标注团队”，进入“标注团队”管理页面。

2. 在“标注团队”管理页面，从左侧团队列表中选择一个团队，单击团队，其右侧区域将呈现“团队详情”。
3. 在“团队详情”区域，单击“添加成员”。
4. 在弹出的“添加成员”对话框中，填写成员的“邮箱”、“描述”、指定“角色”，然后单击“确定”。

   邮箱作为团队管理中的唯一标识，不同成员不能使用同一个邮箱。您填写的邮箱地址将被记录并保存在ModelArts中，仅用于ModelArts团队标注功能，当成员删除后，其填写的邮箱信息也将被一并删除。

   其中，“角色”支持“Labeler”、“Reviewer”和“Team Manager”，“Team Manager”只能设置为一个人。

   图 2-81 添加成员

   成员添加完成后，团队详情区域中将呈现此成员的信息。

修改成员信息

团队中的成员，当其信息发生变化时，可以编辑其基本情况。

1. 在“团队详情”区域，选择需修改的成员。
2. 在成员所在行的“操作”列，单击“修改”。在弹出的对话框中，修改其“描述”或“角色”。

   成员的“邮箱”无法修改，如果需要修改邮箱地址，建议先删除此成员，然后再基于新的邮箱地址添加新成员。

   “角色”支持“Labeler”、“Reviewer”和“Team Manager”，“Team Manager”只能设置为一个人。

删除成员

- **删除单个成员**

  在“团队详情”区域，选择需要删除的成员， 单击“操作”列的“删除”。在弹出的对话框中，确认信息无误后，单击“确定”完成删除操作。

- **批量删除**

  在“团队详情”区域，勾选需删除的成员，然后单击“删除”。在弹出的对话框中，确认信息无误后，单击“确定”完成多个成员的删除操作。

  图 2-82 批量删除

2.14.4 管理团队标注任务

针对启用团队标注功能的数据集，支持创建团队标注任务，将标注任务指派给不同的团队，由多人完成标注任务。同时，在成员进行数据标注过程中，支持发起验收、继续验收以及查看验收报告等功能。

创建团队标注任务

如果您在创建数据集时，即启用团队标注，且指派了某一团队负责标注，系统将默认基于此团队创建一个标注任务。您可以在数据集创建后，在数据集的“标注任务进展”页面查看此任务。

您还可以重新创建一个团队标注任务，指派给同一团队的不同成员，或者指派给其他标注团队。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“数据管理 >数据集”，打开数据集列表。
2. 在数据集列表中，选择支持团队标注的数据集，单击数据集名称进入数据集概览页。
3. 在数据集概览页中，单击“标注任务进展”页签，可查看此数据集已有的标注任务。单击右上角的“创建团队标注任务”开始创建新任务。

   图 2-83 标注任务

4. 在弹出的“创建团队标注任务”对话框中，填写相关参数，然后单击“确定”，完成任务创建。
   - “名称”：设置此任务的名称。
   - “类型”：设置任务类型，支持“指定标注团队”或“指定标注管理员”。
   - “选择标注团队”：任务类型设置为“指定标注团队”，需在此参数中指定一个团队，同时勾选此团队中某几个成员负责标注。下拉框中将罗列当前帐号下创建的标注团队及其成员，团队管理的操作指导请参见团队标注简介。

- “选择标注接口人”：任务类型设置为“指定标注管理员”，需在所有团队的“Team Manager”中选择一人作为管理员。
- “标签集”：展示当前数据集已有的标签及标签属性。在“标签集”下方也可以设置“自动将新增图片同步给标注团队”或“团队标注的图片自动加载智能标注结果”。

说明

团队标注加载智能标注结果的处理步骤：

- 如果类型选择"指定标注团队"，需要先创建团队标注任务，然后执行智能标注任务。
- 如果类型选择"指定标注管理员"，在“我参与的”页签下选择团队标注任务，单击"分配任务"。

图 2-84 创建团队标注任务

任务创建完成后，您可以在“标注任务进展”页签下看到新建的任务。

进入标注（团队成员）

在标注任务创建后，被分配任务的团队成员将收到一封通知邮件，标题为“您有新的标注任务待查收”。

在邮件详情中，单击标注任务链接进入ModelArts管理控制台“数据管理>数据标注”“我参与的”页签下，选择标注任务，完成标注。

不同类型的数据集，标注方式不同，详细请参见：

- 图像分类
- 物体检测
- 文本分类
- 命名实体
- 文本三元组

在标注页面中，每个成员可查看“未标注”、“待确认”、“已驳回”、“待审核”、“审核通过”、“验收通过”的图片信息。请及时关注管理员驳回以及待修正的图片。

当团队标注任务中，分配了Reviewer角色，则需要对标注结果进行审核，审核完成后，再提交给管理员验收。

图 2-85 成员标注平台

任务验收（管理员）

- **发起验收**

  当团队的成员已完成数据标注，数据集的创建者可发起验收，对标注结果进行抽验。只有当标注成员存在标注完成的数据时，才可以发起验收，否则发起验收按钮为灰色。

  a. 在“标注任务进展”页签中，针对需发起验收的任务，单击“发起验收”。

  b. 在弹出的对话框中，设置“抽样策略”，可设置为“按百分比”，也可以设置为“按数量”。设置好参数值后，单击“确定”启动验收。

     “按百分比”：按待验收图片总数的一定比例进行抽样验收。

     “按数量”：按一定数量进行抽样验收。

图 2-86 发起验收

c. 验收启动后，界面将展示实时验收报告，您可以在右侧选择“验收结果”（“通过”或“不通过”）。

当选择验收结果为“通过”时，需设置“验收评分”（分“A”、“B”、“C”、“D”四个选项，“A”表示最高分），如图2-88所示。当选择验收结果为“不通过”时，可以在文本框中写明驳回原因，如图2-89所示。

图 2-87 查看实时验收报告

图 2-88 设置验收结果为“通过”

图 2-89 设置验收结果为“不通过”

- **继续验收**

  针对未完成验收的任务，可以继续验收。针对未发起过验收流程的任务，不支持“继续验收”，按钮为灰色。

  在“标注任务进展”页签中，针对需继续验收的任务，单击“继续验收”。系统直接进入“实时验收报告”页面，您可以继续验收未验收的图片，设置其“验收结果”。

- **完成验收**

  在完成验收窗口，您可以查看本数据集的验收情况，如抽样文件数等，同时设置如下参数，然后进行验收。只有完成验收，标注信息才会同步到数据集的已标注页面中。

  一旦标注数据完成验收，团队成员无法再修改标注信息，只有数据集创建者可修改。

表 2-32 完成验收的参数设置

参数	说明
对已标注数据修改	• 不覆盖：针对同一个数据，不使用当前团队标注的结果覆盖已有数据。 • 覆盖：针对同一个数据，使用当前团队标注的结果覆盖已有数据。覆盖后无法恢复，请谨慎操作。
通过范围	• 全部：当前团队标注完成的所有数据。包含验收通过、未验收和验收不通过的。即本数据集的所有抽样文件数。 • 全部不通过：当前团队标注完成的所有数据不通过验收，即将所有标注数据驳回给标注人员。全部数据指验收通过、未验收和验收不通过的所有数据，即本数据集的所有抽样文件数。 • 验收通过和未验收的数据：针对抽样文件中验收通过和未验收的数据，通过验收。验收不通过的数据将驳回给标注人员。 • 验收通过的数据：针对抽样文件中验收通过的数据，通过验收。未验收和验收不通过的数据将驳回给标注人员。

图 2-90 完成验收

查看验收报告

针对进行中或已完成的标注任务，都可以查看其验收报告。在“标注任务进展”页签中，单击“验收报告”，即可在弹出的“验收报告”对话框中查看详情。

图 2-91 查看验收报告

删除标注任务

针对不再使用的标注任务，您可以在“标注任务进展”页签下，单击任务所在行的删除。任务删除后，未验收的标注详情将丢失，请谨慎操作。但是数据集中的原始数据以及完成验收的标注数据仍然存储在对应的OBS桶中。

2.15 数据处理

2.15.1 数据处理简介

ModelArts平台提供的数据处理功能，基本目的是从大量的、杂乱无章的、难以理解的数据中抽取或者生成对某些特定的人们来说是有价值、有意义的数据。当数据采集和接入之后，数据一般是不能直接满足训练要求的。为了保障数据质量，以免对后续操作（如数据标注、模型训练等）带来负面影响，开发过程通常需要进行数据处理。常见的数据处理类型有以下四种：

- **数据校验**：通常数据采集后需要进行校验，保证数据合法。

  数据校验是指对数据可用性的基本判断和验证的过程。通常，我们采集的数据或多或少都会有很多格式问题，无法被进一步处理。以图像识别为例，用户经常会从网上找一些图片用于训练，但是其质量难以保证，有可能图片的名字、路径、后缀名都不满足训练算法的要求；图片也可能有部分损坏，造成无法解码、无法被算法处理的情况。因此，数据校验非常重要，可以帮助人工智能开发者提前发现数据问题，有效防止数据噪声造成的算法精度下降或者训练失败问题。

- **数据清洗**：数据清洗是指对数据进行去噪、纠错或补全的过程。

数据清洗是在数据校验的基础上，对数据进行一致性检查，处理一些无效值。例如在深度学习领域，可以根据用户输入的正样本和负样本，对数据进行清洗，保留用户想要的类别，去除用户不想要的类别。

- **数据选择**：数据选择一般是指从全量数据中选择数据子集的过程。

  数据可以通过相似度或者深度学习算法进行选择。数据选择可以避免人工采集图片过程中引入的重复图片、相似图片等问题；在一批输入旧模型的推理数据中，通过内置规则的数据选择可以进一步提升旧模型精度。

- 数据增强：

  数据扩增通过简单的数据扩增例如缩放、裁剪、变换、合成等操作直接或间接的方式增加数据量。

  图像生成应用相关深度学习模型，通过对原数据集进行学习，训练生成新的数据集的方式增加数据量。

2.15.2 创建数据处理任务

您可以创建一个数据处理任务，对已有的数据进行数据校验、数据清洗、数据选择或者数据增强操作。

前提条件

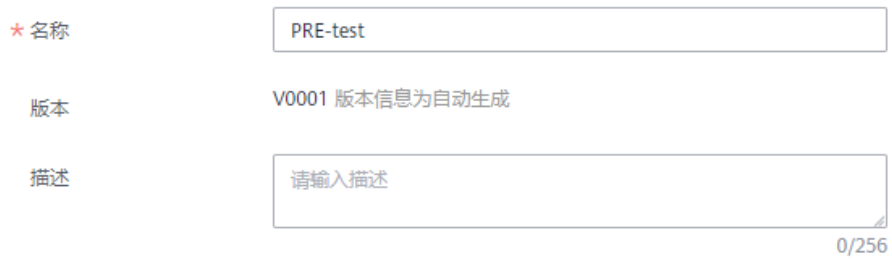
- 数据已准备完成：已经创建数据集或者已经将数据上传至OBS
- 确保您使用的OBS与ModelArts在同一区域

创建数据处理任务

1. 登录ModelArts管理控制台，在左侧的导航栏中选择“数据管理>数据处理”，进入“数据处理”页面。
2. 在“数据处理”页面，单击“创建”进入“创建数据处理”页面。
3. 在创建数据处理页面，填写相关算法参数。
   a. 填写基本信息。基本信息包括“名称”、“版本”和“描述”。其中“版本”信息由系统自动生成，按“V0001”、“V0002”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。

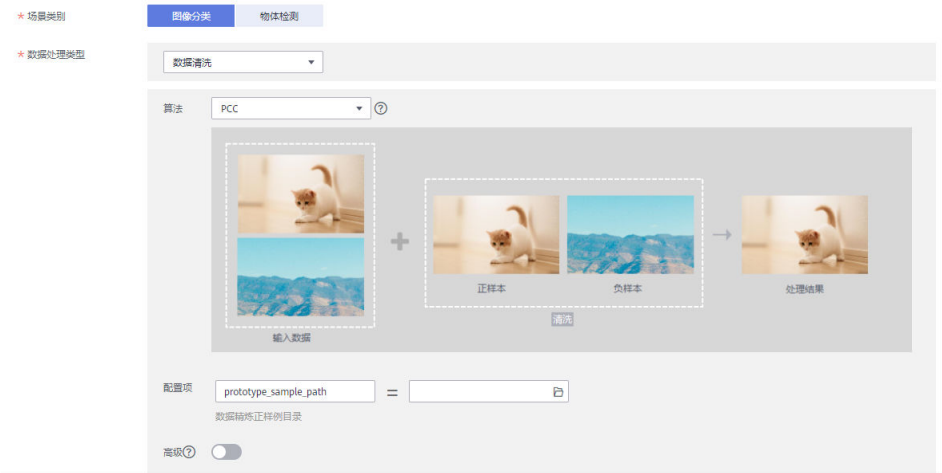
      图 2-92 创建数据处理基本信息

   b. 设置场景类别。场景类别当前支持“图像分类”和“物体检测”。
   c. 设置数据处理类型。数据处理类型支持“数据清洗”、“数据校验”、“数据选择”和“数据增强”。

      针对不同的数据处理类型，您需要填写相应算子的设置参数，算子的详细参数参见**预置算子说明**。

图 2-93 设置场景类别和数据处理类型



d. 设置输入与输出。需根据实际数据情况选择“数据集”或“OBS目录”。设置为“数据集”时，需填写“数据集名称”和“数据集版本”；设置为“OBS目录”时，需填写正确的OBS路径。

图 2-94 输入输出设置-数据集

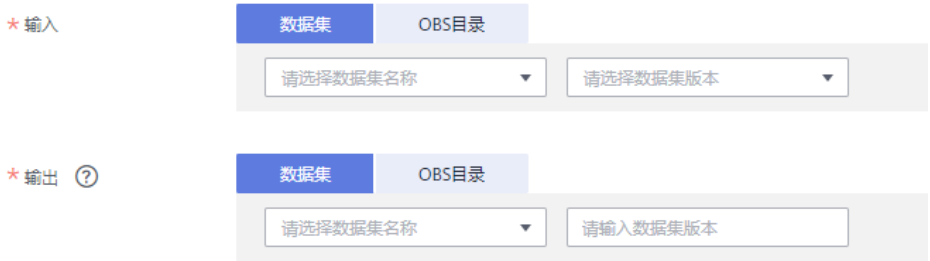


图 2-95 输入输出设置-OBS 目录

e. 确认参数填写无误后，单击“创建”，完成数据处理任务的创建。

2.15.3 管理和查看数据处理任务

删除数据处理任务

当已有的数据处理任务不再使用时，您可以删除数据处理任务。

处于“完成”、“失败”、“已停止”、“运行失败”、“部署中”状态的训练作业，您可以单击操作列的“删除”，删除对应的数据处理任务。

查看数据处理任务详情

1. 登录ModelArts管理控制台，在左侧的导航栏中选择“数据管理>数据处理”，进入“数据处理”页面。
2. 在数据处理列表中，单击数据处理任务名称，进入数据处理任务的版本管理页面。您可以在该页面进行数据处理任务的“修改”与“删除”。

图 2-96 数据处理版本管理页面

3. 您可以在版本管理页面，通过切换页签查看“配置信息”、“日志”和“结果展示”。

图 2-97 日志页面

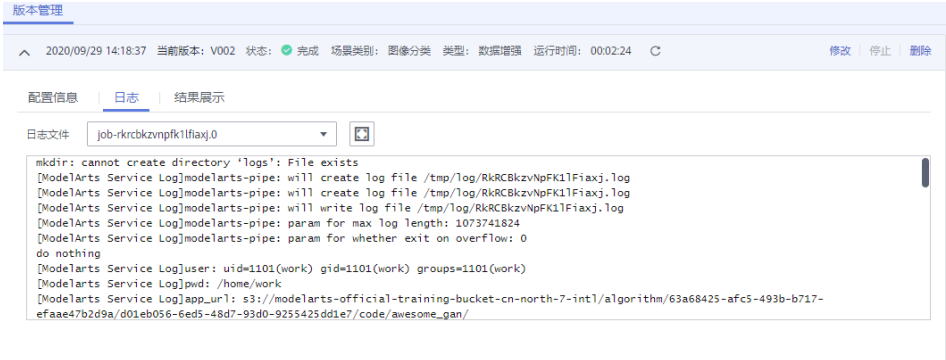
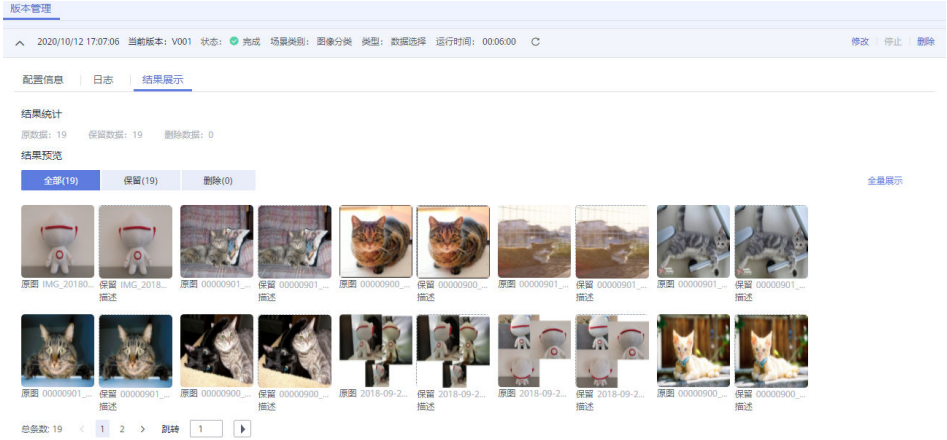


图 2-98 结果展示页面



2.15.4 预置算子说明

2.15.4.1 数据校验

MetaValidation 算子概述

ModelArts的数据校验通过MetaValidation算子实现。当前ModelArts支持jpg、jpeg、bmp、png四种图片格式。物体检测场景支持xml标注格式，不支持“非矩形框”标注。针对您提供的数据集，MetaValidation算子支持对图片和xml文件进行数据校验：

表 2-33 图片类数据校验

异常情况	处理方案
图片本身损坏无法解码	过滤掉不能解码的图片
图片通道可能是1通道、2通道，不是常用的3通道	转换图片成RGB三通道
图片格式不在ModelArts支持的格式范围内	转换图片格式至jpg格式
图片后缀与实际格式不符，但格式在MA支持的格式内	后缀转换成与实际格式一致
图片后缀与实际格式不符，且格式不在MA支持的格式内	转换图片格式至jpg格式
图片分辨率过大	宽、高按指定大小同比例进行裁剪

表 2-34 标注类文件数据校验

异常情况	处理方案
xml结构残缺，无法解析	过滤xml文件

异常情况	处理方案
xml中没有标注“object”	过滤xml文件
xml中没有矩形框“bndbox”	过滤xml文件
某些标注“object”中没有矩形框“bndbox”	过滤标注“object”
图片经过裁剪后，xml文件中宽高不符	修改错误宽高参数为图片真实宽高
xml中没有“width”、“height”字段	根据图片真实宽高补全xml中的“width”、“height”字段和值
图片经过裁剪后，xml中矩形框“bndbox”大小不符	按图片裁剪比例缩放xml文件中“bnxbox”值
xml中矩形框“bndbox”宽或高值过小，显示为一条线	矩形框宽或高差值小于2，移除当前“object”
xml中矩形框“bndbox”最小值大于最大值	移除当前“object”
矩形框“bndbox”超出图片边界，且超出部分占框面积50%以上	移除当前“object”
矩形框“bndbox”超出图片边界，但超出部分小于框面积50%	矩形框“bndbox”拉回到图片边界

说明

数据校验过程不会改动原始数据，通过校验的图片或xml文件保存在指定的输出路径下。

参数说明

表 2-35 数据校验-MetaValidation 算子参数说明

参数名	是否必选	默认值	参数说明
image_max_width	否	-1	输入图片宽度最大值，若输入图片宽度超过设定值则按比例裁剪。单位为px。 默认值 -1 表示不做裁剪。
image_max_height	否	-1	输入图片长度最大值，若输入图片长度超过设定值则按比例裁剪。单位为px。 默认值 -1 表示不做裁剪。

输入要求

算子输入分为两种，“数据集”或“OBS目录”。

- 选择“数据集”，请从下拉框中选择ModelArts中管理的数据集及其版本。要求数据集类型与您在本任务中选择的场景类别一致。
- 选择“OBS目录”，存放结构又分两种情况，“仅包含图片”或“包含图片和标注信息”。
  - “仅包含图片”：当目录下全是图片时，支持jpg、jpeg、png、bmp格式，嵌套子目录的图片也将全部读入。
  - “包含图片和标注信息”：根据不同场景类型，结构不同。

    图像分类场景，其目录结构如下所示。如下目录结构，仅支持单标签场景。

    ```
    input_path/
      --label1/
          ----1.jpg
      --label2/
          ----2.jpg
      --../
    ```

    物体检测场景，其目录结构如下所示。支持jpg、jpeg、png、bmp格式的图片，xml为标准的PACAL VOC格式标注文件。

    ```
    input_path/
      --1.jpg
      --1.xml
      --2.jpg
      --2.xml
      ...
    ```

输出说明

- **图像分类**

  输出数据的目录结构如下所示。

  ```
  output_path/
    --Data/
      ----class1/  # 若输入数据有标注信息会一并输出，class1为标注类别
          ------1.jpg
          ------2_checked.jpg
      ----class2/
          ------3.jpg
          ------4_checked.jpg
      ----5_checked.jpg
    --output.manifest
  ```

  其中manifest文件内容示例如下所示。会给每一条数据加上一个校验属性"property":{"@modelarts:data_checked":true}

  ```
  {
    "id": "xss",
    "source": "obs://hard_example_path/Data/fc8e2688015d4a1784dcbda44d840307_14_checked.jpg",
    "property": {
      "@modelarts:data_checked": true
    },
    "usage": "train",
    "annotation": [
      {
        "name": "Cat",
        "type": "modelarts/image_classification"
      }
    ]
  }
  ```

- **物体检测**

  在输出目录下，文件结构如下所示。

  ```
  output_path/
    --Data/
      ----1_checked.jpg
  ```

```
      ----1_checked.xml  # 若输入数据在校验过程中经过了转换，文件名会加上'_checked'
      ----2.jpg          # 若输入数据未经过转换，则以原来的名字保存
      ----2.xml
  --output.manifest
```

其中manifest文件内容示例如下所示。会给每一条数据加上一个校验属性"property":{"@modelarts:data_checked":true}

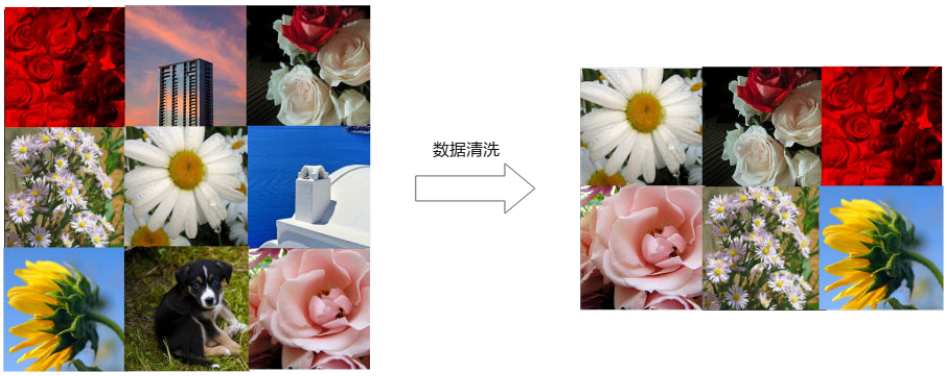
```
{
  "source": "obs://hard_example_path/Data/be462ea9c5abc09f_checked.jpg",
  "property": {
    "@modelarts:data_checked": true
  },
  "annotation": [
    {
      "annotation-loc": "obs://hard_example_path/Data/be462ea9c5abc09f_checked.xml",
      "type": "modelarts/object_detection",
      "annotation-format": "PASCAL VOC",
      "annotated-by": "modelarts/hard_example_algo"
    }
  ]
}
```

2.15.4.2 数据清洗

PCC 算子概述

ModelArts的数据清洗通过PCC算子实现。图像分类或者物体检测的数据集中可能存在非所需类别的图像，需要将这些图像去除掉，以免对标注、模型训练造成干扰。

图 2-99 PCC 算子效果



参数说明

表 2-36 数据清洗-PCC 算子参数说明

参数名	是否必选	默认值	参数说明
prototype_sample_path	是	None	数据清洗正样例目录。目录应存放正样例图片文件，算法将这些图片为正样例，对输入中的数据进行过滤，即保留与“prototype_sample_path”目录下图片相似度高的数据。 请输入一个真实存在的OBS目录，且目录下已包含提供的正样例图片，且以obs://开头。如：*obs://obs_bucket_name/folder_name*
criticism_sample_path	否	None	数据清洗负样例目录。目录应存放负样例图片文件，算法将这些图片为负样例，对算法输入中的数据进行过滤， 即保留与“criticism_sample_path”目录下图片相似度差距较大的数据。 建议该参数和“prototype_sample_path”配合使用，可以提高数据清洗的准确性。 请输入一个真实存在的OBS目录，且以obs://开头。如：*obs://obs_bucket_name/folder_name*
n_clusters	否	auto	数据样本的种类数，默认值auto。您可以输入小于样本总数的整数或auto。auto表示使用正样本目录的图片个数作为数据样本的种类数。
simlarity_threshold	否	0.9	相似度阈值。两张图片相似程度超过阈值时，判定为相似图片，反之按非相似图片处理。输入取值范围为0~1。
embedding_distance	否	0.2	样本特征间距。两张图片样本特征间距小于设定值，判定为相似图片，反之按非相似图片处理。输入取值范围为0~1。
do_validation	否	True	是否做数据校验，可填True或者False。表示数据清洗前需要做数据校验，否则只做数据清洗。

输入要求

算子输入分为两种，“数据集”或“OBS目录”。

- 选择“数据集”，请从下拉框中选择ModelArts中管理的数据集及其版本。要求数据集类型与您在本任务中选择的场景类别一致。
- 选择“OBS目录”，存放结构又分两种情况，“仅包含图片”或“包含图片和标注信息”。
  - “仅包含图片”：当目录下全是图片时，支持jpg、jpeg、png、bmp格式，嵌套子目录的图片也将全部读入。

– “包含图片和标注信息”：根据不同场景类型，结构不同。

图像分类场景，其目录结构如下所示。如下目录结构，仅支持单标签场景。

```
input_path/
    --label1/
        ----1.jpg
    --label2/
        ----2.jpg
    --../
```

物体检测场景，其目录结构如下所示。支持jpg、jpeg、png、bmp格式的图片，xml为标准的PACAL VOC格式标注文件。

```
input_path/
    --1.jpg
    --1.xml
    --2.jpg
    --2.xml
    ...
```

输出说明

- **图像分类**

  输出数据的目录结构如下所示。

```
output_path/
    --Data/
        ----class1/  # 若输入数据有标注信息会一并输出，class1为标注类别
            ------1.jpg
        ----class2/
            ------2.jpg
        ----3.jpg
    --output.manifest
```

  其中manifest文件内容示例如下所示。

```
{
    "id": "xss",
    "source": "obs://home/fc8e2688015d4a1784dcbda44d840307_14.jpg",
    "usage": "train",
    "annotation": [
        {
            "name": "Cat",
            "type": "modelarts/image_classification"
        }
    ]
}
```

- **物体检测**

  输出数据的目录结构如下所示。

```
output_path/
    --Data/
        ----1.jpg
        ----1.xml  # 若输入数据有标注信息会一并输出，xml为标注文件
        ----2.jpg
        ----3.jpg
    --output.manifest
```

  其中manifest文件内容示例如下所示。

```
{
    "source":"obs://fake/be462ea9c5abc09f.jpg",
    "annotation":[
        {
            "annotation-loc":"obs://fake/be462ea9c5abc09f.xml",
            "type":"modelarts/object_detection",
            "annotation-format":"PASCAL VOC",
            "annotated-by":"modelarts/hard_example_algo"
        }
```

```
  ]
}
```

2.15.4.3 数据选择

数据选择算子概述

ModelArts提供以下数据选择算子：

- SimDeduplication：可以依据用户设置的相似程度阈值完成图像去重处理。图像去重是图像数据处理常见的数据处理方法。图像重复指图像内容完全一样，或者有少量的尺度、位移、色彩、亮度变化，或者是添加了少量其他内容等。

图 2-100 SimDeduplication 效果图

表 2-37 高级参数说明

参数名	是否必选	默认值	参数说明
simlarity_threshold	否	0.9	相似程度阈值，两张图片间的相似度大于阈值时，其中一张会作为重复图片被过滤掉。取值范围为0~1。
do_validation	否	True	是否做数据校验，可填True或者False。表示数据去重前需要做数据校验，否则只做数据去重。

- RRD：可以依据用户设置的比例去除差异最大的数据。

图 2-101 RRD 效果图

表 2-38 高级参数说明

参数名	是否必选	默认值	参数说明
sample_ratio	否	0.9	数据留下的百分比。取值范围为0~1。例如0.9表示保留百分之90的原数据。
n_clusters	auto	auto	数据样本的种类数，默认为auto，即按照目录中图片个数取类别总数，可指定具体类别数，如 4
do_validation	否	True	是否做数据校验，可填True或者False。表示数据去冗余前需要做数据校验，否则只做数据去重。

输入要求

算子输入分为两种，“数据集”或“OBS目录”。

- 选择“数据集”，请从下拉框中选择ModelArts中管理的数据集及其版本。要求数据集类型与您在本任务中选择的场景类别一致。
- 选择“OBS目录”，存放结构又分两种情况，“仅包含图片”或“包含图片和标注信息”。
  - “仅包含图片”：当目录下全是图片时，支持jpg、jpeg、png、bmp格式，嵌套子目录的图片也将全部读入。
  - “包含图片和标注信息”：根据不同数据类型，结构不同。

    图像分类，其目录结构如下所示。如下目录结构，仅支持单标签场景。

    ```
    input_path/
      --label1/
          ----1.jpg
      --label2/
          ----2.jpg
      --../
    ```

    物体检测，其目录结构如下所示。支持jpg、jpeg、png、bmp格式的图片，xml为标准的PACAL VOC格式标注文件。

    ```
    input_path/
      --1.jpg
      --1.xml
      --2.jpg
      --2.xml
      ...
    ```

输出说明

- **图像分类**

  输出数据的目录结构如下所示。

  ```
  output_path/
    --Data/
        ----class1/  # 若输入数据有标注信息会一并输出，class1为标注类别
            ------1.jpg
        ----class2/
            ------2.jpg
            ------3.jpg
    --output.manifest
  ```

其中manifest文件内容示例如下所示。

```
{
    "id": "xss",
    "source": "obs://home/fc8e2688015d4a1784dcbda44d840307_14.jpg",
    "usage": "train",
    "annotation": [
        {
            "name": "Cat",
            "type": "modelarts/image_classification"
        }
    ]
}
```

- **物体检测**

  输出数据的目录结构如下所示。

```
output_path/
    --Data/
        ----1.jpg
        ----1.xml  # 若输入数据有标注信息会一并输出，xml为标注文件
        ----2.jpg
        ----3.jpg
    --output.manifest
```

  其中manifest文件内容示例如下所示。

```
{
    "source":"obs://fake/be462ea9c5abc09f.jpg",
    "annotation":[
        {
            "annotation-loc":"obs://fake/be462ea9c5abc09f.xml",
            "type":"modelarts/object_detection",
            "annotation-format":"PASCAL VOC",
            "annotated-by":"modelarts/hard_example_algo"
            }
    ]
}
```

2.15.4.4 数据选择（难例）

算法概述

在实际业务场景中，模型维护是一个长期的过程，比如说按照每周、每月进行数据重训练，或者累计数据至一定量时进行定期的重训练。如果将全量的数据用于重训练，需要耗费较大的标注人力和训练耗时。为了提升模型维护效率，可以采用基于难例数据的重训练 。

难例筛选算法对全量数据进行分析并筛选，仅输出全量数据中少部分对于模型维护有价值的数据。基于筛选后的数据进行重训练，可以有效减少标注人力和训练耗时。

难例筛选算法中融合了多种方法，要达到最佳效果，需要根据实际数据选择部分或全部方法，并调整其权重。

参数说明

参数名	是否必选	默认值	参数说明
source_service	Y	inference	难例任务的前置数据来源，目前仅支持inference。此参数不可修改。

参数名	是否必选	默认值	参数说明
filter_func	Y	comprehensive_mining	难例筛选算法设置为“comprehensive_mining”。此参数不可修改。
checkpoint_path	Y	/home/work/user-job-dir/data_filter/resnet_v1_50	用于特征提取的模型目录，当前仅支持基于Imagenet预训练的resnet_v1_50模型，此参数目前暂时不可修改。
model_serving_url	N	None	推理模型路径。得出推理结果的模型文件路径, 即训练作业的输出路径。该模型会用于aug_consistent_mining算法中的数据增强再推理。 请输入一个真实存在的OBS目录，例如：obs://obs_bucket_name/folder_name/
train_data_path	N	None	训练数据集，model_serving_url模型使用的训练数据，需输入数据集版本生成的manifest。 请输入一个真实存在的OBS目录，例如：obs://obs_bucket_name/folder_name/v001.manifest
comprehensive_algo_config	N	clustering_mining:0.2020+aug_consistent_mining:0.4265+feature_distribution_mining:0.0451+sequential_mining:0.425+image_similarity_mining:0.0949+predict_score_mining:0.3900+anomaly_detection_mining:0.2020	使用的算法及其权重，默认使用系统实验后效果最佳参数，也可以根据不同的数据自行配置。 例如：predict_score_mining:0.3900+anomaly_detection_mining:0.2020
algo_hard_threshold	N	0.1	筛选系数的阈值设置。取值范围0~1。 由于过高阈值可能导致输出结果为0，建议合理填写。
aug_op_config	N	crop:0.1+fliplr:0.1+gaussianblur:0.1	aug_consistent_mining算法中用到的增强手段，支持crop、fliplr、gaussianblur、flipud、scale、translate、shear、superpixels、sharpen、add、invert。

参数名	是否必选	默认值	参数说明
feature_op_config	N	image_aspect_ratio:0.5+image_brightness:1.0+image_saturation:0.5+image_resolution:0.5+image_colorfulness:0.5+ambiguity:1.0+bbox_num:1.0+bbox_iou:1.0+bbox_std:0.5+bbox_bright:0.5+bbox_ambiguity:0.5+bbox_aspect_ratio:1.0+bbox_area_ratio:0.5+bbox_edge_value:0.5	feature_distribution_mining算法中定义的特征，可自行修改权重。
score_threshold_up	N	0.6	predict_score_mining算法中定义的置信度最高值。取值范围0~1。
score_threshold_low	N	0.3	predict_score_mining算法中定义的置信度最低值。取值范围0~1。
margin	N	0.8	top2置信度差值定义, 差值超过该阈值，则为难例。取值范围0~1，默认值为0.8。
similarity_sample_ratio	N	1.0	image_similarity_mining算法中的相似比例定义。取值范围0~1，默认值为1.0。
task_summary_file	N	None	算法简化的日志输出路径及日志文件。请填写一个真实存在的OBS目录，以“obs://”开头。文件名称可自行定义。 例如：obs://obs_bucket_name/folder_name/xxx.log
output_dataset_type	N	manifest	支持directory和manifest类型。 ● directory：将原始图片和标签输出到结果目录的Data文件夹下。 ● manifest：仅输出manifest文件。 该参数在数据处理模块将根据页面选择自动填充。

输入要求

算子输入分为两种，“数据集”或“OBS目录”。

- 选择“数据集”，请从下拉框中选择ModelArts中管理的数据集及其版本。要求数据集类型与您在本任务中选择的场景类别一致。
- 选择“OBS目录”，目录下需包含用于推理的原始图片以及推理结果文件

  其目录结构如下所示。

```
input_path/
   --images/ # 文件夹名称必须为images
       ----1.jpg
       ----2.jpg
   --inference_results/  # 文件夹名称必须为inference_results
       ----1.jpg_result.txt
       ----2.jpg_result.txt
```

  其中，推理结果文件txt的内容需满足如下格式。如果使用ModelArts预置算法训练得到的模型进行推理，默认的推理结果已满足要求。

  - **图像分类**

```
{
   "predicted_label": "dog",
   "scores": [
      [
         "dog",
         "0.589"
      ],
      [
         "cat",
         "0.411"
      ]
   ]
}
```

  - **物体检测**

```
{
   "detection_classes": [
      "cat",
      "cat"
      ],
   "detection_boxes": [
      [
         117.56356048583984,
         335.9902648925781,
         270.50848388671875,
         469.0136413574219
      ],
      [
         18.747316360473633,
         13.10757064819336,
         217.25146484375,
         108.3551025390625
      ]
   ],
   "detection_scores": [
         0.5179755091667175,
         0.46941104531288147
      ]
}
```

输出说明

- 目标检测

  输出目录结构如下所示。

```
output_path :
   --Data
```

```
        ----1.jpg
        ----1.xml      # 导出筛选结果至该目录下
    --output.manifest
```

其中manifest文件示例如下所示。

```
{"source":"/tmp/test_out/object_detection/images/be462ea9c5abc09f.jpg",
"hard":"True",
"hard-reasons":"0",  # 判定该样本为难例的原因，具体原因目前只在智能标注模块展示
"hard-coefficient":"1.0", # 难例算法得到的难例系数，越大代表可能是难例的概率越大
"annotation":[
{"annotation-loc":"/tmp/test_out/object_detection/annotations/be462ea9c5abc09f.xml",
"type":"modelarts/object_detection",
"annotation-format":"PASCAL VOC",
"annotated-by":"modelarts/hard_example_algo"}]}
```

- 图像分类

  输出目录结构如下所示。

```
output_path :
    --Data
        ----class1
            ------1.jpg
        ----class2
            ------2.jpg
    --output.manifest
```

其中manifest文件示例如下所示。

```
{"source":"obs://obs_bucket_name/folder_name/catDog/5.jpg",
"hard":true,
"hard-reasons":"1-20-2-19-21-3",
"hard-coefficient":1.0,
"annotation":[
{"name":"cat",
"type":"modelarts/image_classification",
"confidence":0.599,
"annotated-by":"modelarts/hard_example_algo"}]}
```

日志文件说明

task_summary_file 为简化日志的输出文件路径，内容如下：

```
{
"task_status": 'SUCCEED', # 算法执行状态
"total_sample": integer,  # 输入样本总数
"hard_sample": integer    # 输出样本总数
}
```

或者是

```
{
"task_status": 'FAILED',
"error_message": 'xxxxxx' # 导致算法执行失败的异常信息
}
```

2.15.4.5 数据增强（数据扩增）

数据扩增算子概述

数据扩增主要用于训练数据集不足或需要仿真的场景，能通过对已标注的数据集做变换操作来增加训练图片的数量，同时会生成相应的标签。在深度学习领域，增强有重要的意义，能提升模型的泛化能力，增加抗扰动的能力。数据扩增过程不会改动原始数据，扩增后的图片或xml文件保存在指定的输出路径下。

ModelArts提供以下数据扩增算子：

表 2-39 数据扩增算子介绍

算子	算子说明	高级
AddNoise	添加噪声，模拟常见采集设备在采集图片过程中可能会产生的噪声。	● noise_type：添加噪声的分布类型，Gauss为高斯噪声，Laplace为拉普拉斯噪声，Poisson是泊松噪声，Impulse是脉冲噪声，SaltAndPepper为椒盐噪声。默认值为Gauss ● loc：噪声分布的均值，仅在Gauss和Laplace生效。默认值为0 ● scale：噪声分布的标准差，仅在Gauss和Laplance生效。默认值为1 ● lam：泊松分布的lambda系数，仅在Poisson有效。默认值为2 ● p：对于每个像素点，出现脉冲噪声或椒盐噪声的概率，仅在Impulse和SaltAndPepper有效。默认值为0.01 ● do_validation：数据扩增前是否做数据校验。默认值为True。
Blur	模糊，使用滤波器对图像进行滤波操作，有时用于模拟成像设备的成像。	● blur_type：可选Gauss和Average两种模式，分别为高斯和均值滤波。默认值为Gauss ● do_validation：数据扩增前是否做数据校验。默认值为True。
Crop	图片裁剪，随机裁剪图片的一部分作为新的图片。	● crop_percent_min：各边裁剪占比的随机取值范围的最小值。默认值为0.0 ● crop_percent_max：各边裁剪占比的随机取值范围的最大值。默认值为0.2 ● do_validation：数据扩增前是否做数据校验。默认值为True。
CutOut	随机擦除，在深度学习中常用的方法，用于模拟物体被障碍物遮挡。	do_validation：数据扩增前是否做数据校验。默认值为True。
Flip	翻转，沿图片水平轴或竖直轴做翻转，是非常常见的增强方法。	● lr_ud：选择翻转的方向,lr为水平翻转，ud为竖直翻转。默认值为lr ● flip_p：做翻转操作的概率。默认值为1。 ● do_validation：数据扩增前是否做数据校验。默认值为True。
Grayscale	图片灰度化，将三通道的彩色图像转换到三通道的灰度图像。	do_validation：数据扩增前是否做数据校验。默认值为True。

算子	算子说明	高级
HistogramEqual	直方图均衡化，多半是使用于让图片的视觉效果更加好，在某些场景下会使用。	do_validation：数据扩增前是否做数据校验。默认值为True。
LightArithmetic	亮度增强 ，对亮度空间做线性增强操作。	do_validation：数据扩增前是否做数据校验。默认值为True。
LightContrast	亮度对比度增强，使用一定的非线性函数改变亮度空间的亮度值。	func：默认值为gamma ● gamma为常见方法伽马矫正，公式为255*((v/255)**gamma)') ● sigmoid为函数为S型曲线，公式为255*1/(1+exp(gain*(cutoff-I_ij/255)))') ● log为对数函数，公式为255*gain*log_2(1+v/255) ● linear为线性函数，公式为127 + alpha*(v-127)') do_validation：数据扩增前是否做数据校验。默认值为True。
MotionBlur	运动模糊，模拟物体运动时产生的残影现象。	do_validation：数据扩增前是否做数据校验。默认值为True。
Padding	图片填充，在边缘添加黑色的边。	● px_top：图像顶端增加的像素行数。默认值为1 ● px_right：图像右侧增加的像素行数。默认值为1 ● px_left：图像左侧增加的像素行数。默认值为1 ● px_bottom：图像底侧增加的像素行数。默认值为1 ● do_validation：数据扩增前是否做数据校验。默认值为True。
Resize	调整图片大小。	● height：变换后的图片高度。默认值224 ● width：变换后的图片宽度。默认值224 ● do_validation：数据扩增前是否做数据校验。默认值为True。

算子	算子说明	高级
Rotate	旋转，将图像围绕中心点旋转的操作，操作完成之后保持图片原本的形状不变，不足的部分用黑色填充。	● angle_min：旋转角度随机取值范围的最小值，每张图片会从范围中随机取值作为自己的参数。默认值为90° ● angle_max：旋转角度随机取值范围的最大值，每张图片会从范围中随机取值作为自己的参数。默认值为-90° ● do_validation：数据扩增前是否做数据校验。默认值为True。
Saturation	色度饱和度增强，对图片的HSV中的H和S空间做线性的变化，改变图片的色度和饱和度。	do_validation：数据扩增前是否做数据校验。默认值为True。
Scale	图片缩放，将图片的长或宽随机缩放到一定倍数。	● scaleXY：缩放方向，X为水平，Y为垂直。默认值为X ● scale_min：缩放比例随机取值范围的最小值。默认为0.5 ● scale_max：缩放比例随机取值范围的最大值。默认值为1.5 ● do_validation：数据扩增前是否做数据校验。默认值为True。
Sharpen	图像锐化，用于将边缘清晰化，让物体边缘更加明显。	do_validation：数据扩增前是否做数据校验。默认值为True。
Shear	图片错切，一般用于图片的几何变换，通过线性函数将像素点进行映射。	● shearXY：错切方向，X为水平，Y为竖直。默认值为X ● shear_min：错切角度随机取值范围的最小值。默认值为-30 ● shear_max：错切角度随机取值范围的最大值。默认值为30 ● do_validation：数据扩增前是否做数据校验。默认值为True。
Translate	图片平移，将图片整体像X轴或Y轴平移，超出原图部分舍弃，丢失部分用黑色填充。	● translateXY：平移的方向，X为水平，Y为竖直。默认值为X ● do_validation：数据扩增前是否做数据校验。默认值为True。

算子	算子说明	高级
Weather	添加天气，模拟天气效果。	weather_mode：添加天气的模式，默认值为Rain。 ● Rain：下雨 ● Fog：雾 ● Snow：雪 ● Clouds：云 do_validation：数据扩增前是否做数据校验。默认值为True。

输入要求

算子输入分为两种，“数据集”或“OBS目录”。

- 选择“数据集”，请从下拉框中选择ModelArts中管理的数据集及其版本。要求数据集类型与您在本任务中选择的场景类别一致。
- 选择“OBS目录”，存放结构支持“包含图片和标注信息”模式。

  “包含图片和标注信息”，根据不同场景类型，结构不同。

  图像分类场景，其目录结构如下所示。如下目录结构，仅支持单标签场景。

```
input_path/
    --label1/
        ----1.jpg
    --label2/
        ----2.jpg
    --../
```

  物体检测场景，其目录结构如下所示。支持jpg、jpeg、png、bmp格式的图片，xml为标准的PACAL VOC格式标注文件。

```
input_path/
    --1.jpg
    --1.xml
    --2.jpg
    --2.xml
    ...
```

输出说明

由于算法中有些操作将会舍弃一些数据，输出文件夹里可能不包含全量数据集。例如，“Rotate”会舍弃标注框超出原始图片边界的图片。

输出目录结构如下所示。其中“Data”文件夹用于存放新生成的图片和标注信息，“manifest”文件存储文件夹中图片的结构，可直接导入到数据管理的数据集中。

```
|----data_url
    |----Data
        |----xxx.jpg
        |----xxx.xml(xxx.txt)
    |----output.manifest
```

其中manifest文件内容示例如下所示。

```
{
    "id": "xss",
    "source": "obs://home/fc8e2688015d4a1784dcbda44d840307_14.jpg",
```

```
  "usage": "train",
  "annotation": [
    {
      "name": "Cat",
      "type": "modelarts/image_classification"
    }
  ]
}
```

2.15.4.6 数据增强（图像生成）

图像生成算子概述

图像生成算子利用Gan网络依据已知的数据集生成新的数据集。Gan是一个包含生成器和判别器的网络，生成器从潜在空间中随机取样作为输入，其输出结果需要尽量模仿训练集中的真实样本。判别器的输入则为真实样本或生成网络的输出，其目的是将生成网络的输出从真实样本中尽可能分辨出来。而生成网络则要尽可能地欺骗判别网络。两个网络相互对抗、不断调整参数，最终目的是使判别网络无法判断生成网络的输出结果是否真实。训练中获得的生成器网络可用于生成与输入图片相似的图片，用作新的数据集参与训练。基于Gan网络生成新的数据集不会生成相应的标签。图像生成过程不会改动原始数据，新生成的图片或xml文件保存在指定的输出路径下。

ModelArts提供两种类型的图像生成算子：

- CycleGan算子：基于CycleGAN用于生成域迁移的图像，即将一类图片转换成另一类图片，把X空间中的样本转换成Y空间中的样本。CycleGAN可以利用非成对数据进行训练。模型训练时运行支持两个输入，分别代表数据的原域和目标域，在训练结束时会生成所有原域向目标域迁移的图像。

图 2-102 CycleGan 算子

表 2-40 CycleGan 算子高级参数

参数名	默认值	参数说明
do_validation	True	是否做数据校验，默认为True，表示数据生成前需要做数据校验，否则只做数据生成。
image_channel	3	生成图像的通道数。
image_height	256	图像相关参数：生成图像的高，大小需要是2的次方。
image_width	256	图像相关参数：生成图像的宽，大小需要是2的次方

参数名	默认值	参数说明
batch_size	1	训练相关参数：批量训练样本个数。
max_epoch	100	训练相关参数：训练遍历数据集次数。
g_learning_rate	0.0001	训练相关参数：生成器训练学习率。
d_learning_rate	0.0001	训练相关参数：判别器训练学习率。
log_frequency	5	训练相关参数：日志打印频率（按step计数）。
save_frequency	5	训练相关参数：模型保存频率（按epoch计数）。
predict	False	是否进行推理预测，默认为False。如果设置True，需要在resume参数设置已经训练完成的模型的obs路径。
resume	empty	如果predict设置为True，需要填写Tensorflow模型文件的obs路径用于推理预测。当前仅支持“.pb”格式的模型。示例：obs://xxx/xxxx.pb。 默认值为empty。

- StyleGan算子：基于StyleGan2用于在数据集较小的情形下，随机生成相似图像。StyleGAN提出了一个新的生成器结构，能够控制所生成图像的高层级属性(high-level attributes)，如发型、雀斑等；并且生成的图像在一些评价标准上得分更好。而本算法又增加了数据增强算法，可以在较少样本的情况下也能生成较好的新样本，但是样本数尽量在70张以上，样本太少生成出来的新图像不会有太多的样式。

图 2-103 StyleGan 算子

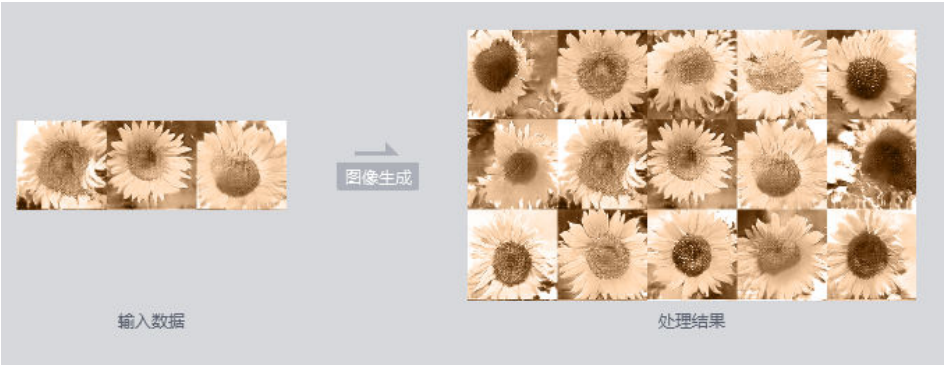


表 2-41 StyleGan 算子高级参数

参数名	默认值	参数说明
resolution	256	生成正方形图像的高宽，大小需要是2的次方。

参数名	默认值	参数说明
batch-size	8	批量训练样本个数。
total-kimg	300	总共训练的图像数量为total_kimg*1000。
generate_num	300	生成的图像数量，如果是多个类的，则为每类生成的数量。
predict	False	是否进行推理预测，默认为False。如果设置True，需要在resume参数设置已经训练完成的模型的obs路径。
resume	empty	如果predict设置为True，需要填写Tensorflow模型文件的obs路径用于推理预测。当前仅支持“.pb”格式的模型。示例：obs://xxx/xxxx.pb。 默认值为empty。
do_validation	True	是否做数据校验，默认为True，表示数据生成前需要做数据校验，否则只做数据生成。

数据输入

算子输入分为两种，“数据集”或“OBS目录”。

- 选择“数据集”，请从下拉框中选择ModelArts中管理的数据集及其版本。要求数据集类型与您在本任务中选择的场景类别一致。
- 选择“OBS目录”，图像生成算子不需要标注信息，输入支持单层级或双层级目录，存放结构支持“单层级”或“双层级”模式。

单层级目录结构如下所示：

```
image_folder----0001.jpg
          ----0002.jpg
          ----0003.jpg
          ...
          ----1000.jpg
```

双层级目录结构如下所示：

```
image_folder----sub_folder_1----0001.jpg
                        ----0002.jpg
                        ----0003.jpg
                        ...
                        ----0500.jpg
          ----sub_folder_2----0001.jpg
                        ----0002.jpg
                        ----0003.jpg
                        ...
                        ----0500.jpg
                        ...
          ----sub_folder_100----0001.jpg
                          ----0002.jpg
                          ----0003.jpg
                          ...
                          ----0500.jpg
```

输出说明

输出目录的结构如下所示。其中“model”文件夹存放用于推理的“frozen pb”模型，“samples”文件夹存放训练过程中输出图像，“Data”文件夹存放训练模型生成的图像。

```
train_url----model----CYcleGan_epoch_10.pb
                 ----CYcleGan_epoch_20.pb
                 ...
                 ----CYcleGan_epoch_1000.pb
        ----samples----0000_0.jpg
                 ----0000_1.jpg
                 ...
                 ----0100_15.jpg
        ----Data----CYcleGan_0_0.jpg
                 ----CYcleGan_0_1.jpg
                 ...
                 ----CYcleGan_16_8.jpg
        ----output_0.manifest
```

其中manifest文件内容示例如下所示。

```
{
    "id": "xss",
    "source": "obs://home/fc8e2688015d4a1784dcbda44d840307_14.jpg",
    "usage": "train",
    "annotation": [
        {
            "name": "Cat",
            "type": "modelarts/image_classification"
        }
    ]
}
```

3 训练管理（旧版即将下线）

3.1 模型训练简介

说明

当前ModelArts同时存在新版训练管理和旧版训练管理功能。旧版训练管理功能仅对部分存量用户可见，新用户不可见。

当前章节仅介绍了旧版训练，旧版训练即将下线，推荐用户使用新版训练，新版训练管理文档请参考模型训练。

新旧训练的差异请参考常见问题。

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用于模型训练。

模型训练功能说明

表 3-1 功能说明

功能	说明	详细指导
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、并且支持查看评估详情。	创建训练作业简介 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和MindInsight是可视化工具，能够有效地展示模型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。	管理可视化作业

3.2 订阅算法

ModelArts的AI Gallery，发布了较多官方算法，可以帮助AI开发者快速开始训练和部署模型。对于不熟悉ModelArts的用户，可以快速订阅官方推荐算法实现模型训练全流程。

AI Gallery不仅可以订阅官方发布算法，也支持用户发布自定义算法和订阅其他开发者分享的算法。为了使用他人或者ModelArts官方分享的算法，您需要将AI Gallery的算法订阅至您的ModelArts中。

- 查找算法
- 订阅算法

查找算法

为了获得匹配您业务的算法，您可以通过多个入口区查找算法。

- 在ModelArts控制台，“算法管理>我的订阅”中，单击“订阅更多算法”，可跳转至“AI Gallery”页面，查找相应的算法。
- 在ModelArts控制台，直接在左侧菜单栏中选择“AI Gallery”，进入“AI Gallery”页面，在“资产集市 > 算法”页面查找相应的算法。

订阅算法

1. 进入“AI Gallery”，选择“资产集市 > 算法”页签，查找您需要的算法并单击算法名称，进入算法详情页，单击算法详情页右侧的“订阅”，订阅您所需的算法。

   订阅后的算法，状态变为“已订阅”，并且将自动展现在“算法管理 > 我的订阅”页面中。

2. 单击“前往控制台”，选择云服务区域，进入“算法管理 > 我的订阅”页面进入此页面内，单击“产品名称”左侧的小三角，展开算法详情，在“版本列表”区域，单击“创建训练作业”即可进行后续操作。

图 3-1 订阅算法

3.3 常用框架

本章详细介绍ModelArts支持的常用AI框架以及使用AI框架编写创建训练作业的训练代码。

训练管理支持的 AI 常用框架

当前ModelArts支持的AI引擎及对应版本如下所示。

表 3-2 旧版训练作业支持的 AI 引擎

工作环境	适配芯片	系统架构	系统版本	AI引擎与版本	支持的cuda或Ascend版本
TensorFlow	CPU/GPU	x86_64	Ubuntu16.04	TF-1.8.0-python3.6	-
				TF-1.13.1-python3.6	-
				tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64	cuda10.1
MXNet	CPU/GPU	x86_64	Ubuntu16.04	MXNet-1.2.1-python3.6	-
Spark_MLlib	CPU	x86_64	Ubuntu16.04	Spark-2.3.2-python3.6	-

工作环境	适配芯片	系统架构	系统版本	AI引擎与版本	支持的cuda或Ascend版本
Ray	CPU/GPU	x86_64	Ubuntu16.04	RAY-0.7.4-python3.6	-
XGBoost-Sklearn	CPU	x86_64	Ubuntu16.04	XGBoost-0.80-Sklearn-0.18.1-python2.7	-
				XGBoost-0.80-Sklearn-0.18.1-python3.6	-
PyTorch	CPU/GPU	x86_64	Ubuntu16.04	PyTorch-1.0.0-python3.6	-
				PyTorch-1.3.0-python3.6	-
				PyTorch-1.4.0-python3.6	-
				pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64	cuda10.2
Ascend-Powered-Engine	Ascend	aarch64	Euler2.8	mindspore_1.7.0-cann_5.1.0-py_3.7-euler_2.8.3-aarch64	5.1.0
				tensorflow_1.15-cann_5.1.0-py_3.7-euler_2.8.3-aarch64	5.1.0
MPI	CPU/GPU	x86_64	Ubuntu18.04	mindspore_1.3.0-cuda_10.1-py_3.7-ubuntu_1804-x86_64	cuda10.1
Caffe	CPU/GPU	x86_64	Ubuntu16.04	Caffe-1.0.0-python2.7	cuda8.0

说明

- MoXing是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**。
- **Ascend-Powered-Engine**仅在“华北-北京四”区域支持。

使用常见框架的训练代码开发

当您使用常用框架创建训练作业时，您需要在创建页面提供代码目录路径、代码目录路径中的启动文件、训练数据路径以及训练输出路径。这四种路径搭建了用户和ModelArts后台交互的桥梁。

- 代码目录路径

  您需要在OBS桶中指定代码目录，并将训练代码、依赖安装包或者预生成模型等训练所需文件上载至该代码目录下。训练作业创建完成后，ModelArts会将代码目录及其子目录下载至后台容器中。

- 代码目录路径中的启动文件

  代码目录路径中的启动文件作为训练启动的入口，当前只支持python格式。

- 训练数据路径

  请注意不要将训练数据路径放在代码目录路径下。训练数据比较大，训练代码目录在启动后会下载至后台，可能会有下载失败的风险。

  在训练作业启动后，ModelArts会挂载硬盘至“/cache”目录，用户可以使用此目录来存储临时文件。“/cache”目录大小请参考**训练环境中不同规格资源“/cache”目录的大小**。

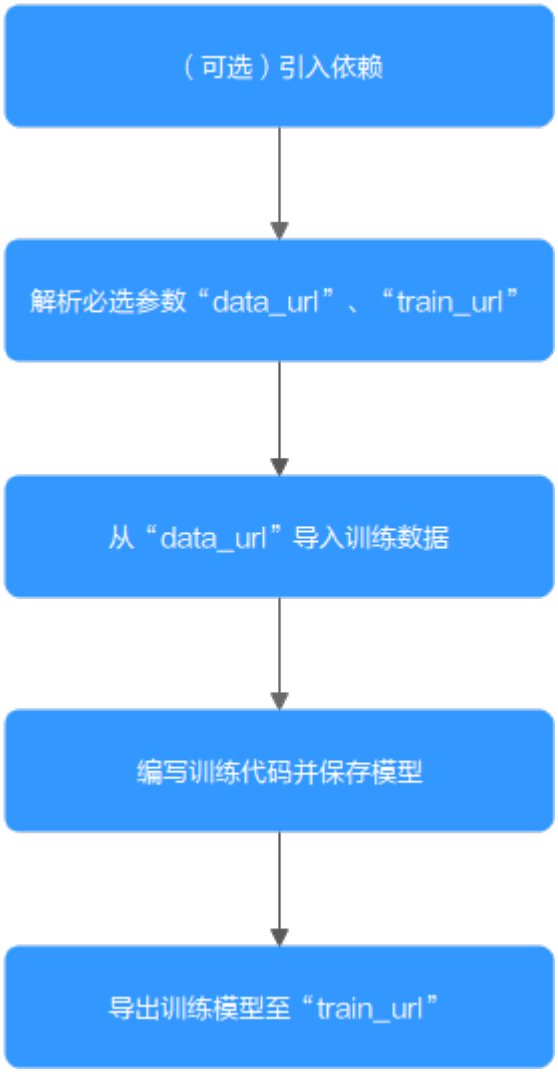
  您需要将训练数据上传至OBS桶另外的路径，并在训练代码中需通过解析命令行参数“data_url”下载训练数据至“/cache”目录。请保证您设置的桶路径有读取权限。

- 训练输出路径

  建议设置一个空目录为训练输出路径。在训练代码中，您需要解析命令行参数“train_url”上载训练输出至指定的训练输出路径，请保证您设置的桶路径有写入权限和读取权限。

当您使用常用框架创建训练作业时，您需要实现训练代码的开发。在ModelArts中，训练代码需包含以下步骤：

图 3-2 训练代码开发说明



1. （可选）引入依赖

   当您使用常见框架创建训练作业的时候，如果您的模型引用了其他依赖，您需要在创建训练作业的“代码目录”下放置相应的文件或安装包。

   - 安装python依赖包请参考**模型中引用依赖包时，如何创建训练作业?**
   - 安装C++的依赖库请参考**如何安装C++的依赖库?**
   - 在预训练模型中加载参数请参考**如何在训练中加载部分训练好的参数?**

   图 3-3 选择常用框架并指定模型启动文件

2. 解析必选参数“data_url”、“train_url”

   在使用常见框架创建训练作业时，您需要在创建训练作业页面填写作业参数配置相关信息。

   “data_url”：训练数据是训练代码开发中必不可少的输入。在创建训练作业时，您需要在作业参数配置“数据来源”。训练代码中的“data_url”指代“数据来源”的路径。

   “train_url”：模型训练结束后，训练模型以及相关输出信息需保存在OBS路径。在创建训练作业时，您需要在作业参数配置“训练输出位置”。训练代码中的“train_url”指代“训练输出位置”的OBS路径。

   图 3-4 作业参数配置相关信息

   在训练代码中需解析“data_url”、“train_url”，ModelArts推荐以下方式实现参数解析。

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

3. 从“data_url”导入训练数据

   已知训练数据路径为“data_url”，ModelArts推荐采用**Moxing接口**实现训练数据下载到“cache”目录。

```
import moxing as mox
mox.file.copy_parallel(args.data_url, “/cache”)
```

4. 训练代码正文和保存模型

   训练代码正文和保存模型涉及的代码与您使用的AI引擎密切相关。以下案例以Tensorflow框架为例，训练代码中解析参数方式采用tensorflow接口tf.flags.FLAGS接受命令行参数：

```
from __future__ import absolute_import
from __future__ import division
from __future__ import print_function

import os

import tensorflow as tf
from tensorflow.examples.tutorials.mnist import input_data

import moxing as mox
```

```
tf.flags.DEFINE_integer('max_steps', 1000, 'number of training iterations.')
tf.flags.DEFINE_string('data_url', '/home/jnn/nfs/mnist', 'dataset directory.')
tf.flags.DEFINE_string('train_url', '/home/jnn/temp/delete', 'saved model directory.')

FLAGS = tf.flags.FLAGS


def main(*args):
    mox.file.copy_parallel(FLAGS.data_url, '/cache/data_url')

    # Train model
    print('Training model...')
    mnist = input_data.read_data_sets('/cache/data_url', one_hot=True)
    sess = tf.InteractiveSession()
    serialized_tf_example = tf.placeholder(tf.string, name='tf_example')
    feature_configs = {'x': tf.FixedLenFeature(shape=[784], dtype=tf.float32),}
    tf_example = tf.parse_example(serialized_tf_example, feature_configs)
    x = tf.identity(tf_example['x'], name='x')
    y_ = tf.placeholder('float', shape=[None, 10])
    w = tf.Variable(tf.zeros([784, 10]))
    b = tf.Variable(tf.zeros([10]))
    sess.run(tf.global_variables_initializer())
    y = tf.nn.softmax(tf.matmul(x, w) + b, name='y')
    cross_entropy = -tf.reduce_sum(y_ * tf.log(y))

    tf.summary.scalar('cross_entropy', cross_entropy)

    train_step = tf.train.GradientDescentOptimizer(0.01).minimize(cross_entropy)

    correct_prediction = tf.equal(tf.argmax(y, 1), tf.argmax(y_, 1))
    accuracy = tf.reduce_mean(tf.cast(correct_prediction, 'float'))
    tf.summary.scalar('accuracy', accuracy)
    merged = tf.summary.merge_all()
    test_writer = tf.summary.FileWriter('/cache/train_url', flush_secs=1)

    for step in range(FLAGS.max_steps):
        batch = mnist.train.next_batch(50)
        train_step.run(feed_dict={x: batch[0], y_: batch[1]})
        if step % 10 == 0:
            summary, acc = sess.run([merged, accuracy], feed_dict={x: mnist.test.images, y_:
mnist.test.labels})
            test_writer.add_summary(summary, step)
            print('training accuracy is:', acc)
    print('Done training!')

    builder = tf.saved_model.builder.SavedModelBuilder(os.path.join('/cache/train_url', 'model'))

    tensor_info_x = tf.saved_model.utils.build_tensor_info(x)
    tensor_info_y = tf.saved_model.utils.build_tensor_info(y)

    prediction_signature = (
        tf.saved_model.signature_def_utils.build_signature_def(
            inputs={'images': tensor_info_x},
            outputs={'scores': tensor_info_y},
            method_name=tf.saved_model.signature_constants.PREDICT_METHOD_NAME))

    builder.add_meta_graph_and_variables(
        sess, [tf.saved_model.tag_constants.SERVING],
        signature_def_map={
            'predict_images':
                prediction_signature,
        },
        main_op=tf.tables_initializer(),
        strip_default_attrs=True)

    builder.save()

    print('Done exporting!')
```

```
    mox.file.copy_parallel('/cache/train_url', FLAGS.train_url)


if __name__ == '__main__':
    tf.app.run(main=main)
```

5. 导出训练模型至“train_url”

   已知训练输出位置为“train_url”，ModelArts推荐采用**Moxing接口**实现输出结果从后台自定义目录“/cache/train_url”目录导出至“train_url”目录。

```
mox.file.copy_parallel(“/cache/train_url”, args.train_url)
```

3.4 创建训练作业

3.4.1 创建训练作业简介

在整个AI全流程开发过程中，ModelArts支持多种类型的训练作业。根据您的算法来源不同，选择不同的创建方式。

训练作业的几种算法来源

- **算法管理**

  算法管理中，管理了用户自己创建的算法和AI Gallery订阅的算法，您可以使用算法管理中的算法，快速创建训练作业，构建模型。操作指导请参见**使用已有算法训练模型**。

- **常用框架**

  如果您已在本地使用一些常用框架完成算法开发，您可以选择常用框架，创建训练作业来构建模型，操作指导请参见**使用常用框架训练模型**。

- **自定义**

  如果您开发算法时使用的框架并不是常用框架，您可以将算法构建为一个自定义镜像，通过自定义镜像创建训练作业。创建训练作业的操作指导请参见**使用自定义镜像训练模型**，自定义镜像的相关规范和说明请参见**训练作业自定义镜像规范**。

3.4.2 使用已有算法训练模型

针对您创建的算法，或者是从AI Gallery订阅的算法，支持快速使用此算法创建训练作业，构建模型。

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上传至OBS目录。
- “算法管理”中，已创建算法或者**订阅算法**。新创建的算法仅在新版训练中支持，请参见**模型开发>创建算法**章节。
- 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，确保帐户未欠费。
- 确保您使用的OBS目录与ModelArts在同一区域。

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“训练管理 > 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，包含“名称”和“描述”。“版本”信息由系统自动生成，按“V0001”、“V0002”规则命名，用户无法修改。
   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3-3**。数据来源的可选范围与已有算法的使用约束保持一致。

表 3-3 作业参数说明

参数名称	子参数	说明
算法来源	算法管理	选择“算法管理”，在“算法名称”右侧单击“选择”，进入“算法管理”对话框。 ● 选择“我的算法”页签，您可以根据实际需求选择已创建成功的算法。新创建的算法仅在新版训练中支持，请参见**模型开发>创建算法**章节。 ● 选择“我的订阅”页签，您可以根据实际需求选择AI Gallery中已订阅的算法。您需要在目标算法的左侧单击下拉三角标，选择合适的版本用于训练作业的创建。如何订阅AI Gallery算法详见**订阅算法**。
训练输入	数据来源>数据集	从ModelArts数据管理中选择可用的数据集及其版本。 ● “选择数据集”：从右侧下拉框中选择ModelArts系统中已发布的数据集。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”指定版本。
	数据来源>数据存储位置	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。
训练输出	模型输出	选择训练结果的存储位置（OBS路径）。为避免出现错误，建议选择一个空目录用作“模型输出”，请勿将数据集存储的目录作为训练输出位置。

参数名称	子参数	说明
超参	-	此参数根据您选择的算法不同而不同。 如果创建的算法或订阅的算法，定义了相关的调优参数，则需在创建训练作业时，填写对应调优参数的参数值。您可以单击“增加超参”，添加多条。
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。训练参数的可选范围与已有算法的使用约束保持一致。

表 3-4 资源参数说明

参数名称	说明
资源池	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”。
规格	针对不同的资源类型，选择资源规格。GPU资源性能更佳，CPU资源性价比更高。如果您的算法已定义使用CPU或GPU，根据已有算法约束条件，您可以在有效规格选择合适的资源规格，无效选项置灰不可选。 Ascend资源仅在“华北-北京四”可用。 不同的资源类型的数据盘容量是不同的，详细介绍参考**训练环境中不同规格资源“/cache”目录的大小**。
计算节点个数	选择计算节点的个数。默认值为“1”。

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 3-5 配置训练作业订阅消息

表 3-5 订阅消息及作业参数的参数说明

参数名称	说明
订阅消息	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。 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。 ● “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，分别代表作业运行中、运行成功、运行失败。
保存作业参数	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。 勾选“保存作业参数”，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。

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“提交”，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”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。

3.4.3 使用常用框架训练模型

如果您在本地使用一些常用框架完成算法开发，如TensorFlow、MindSpore等AI引擎，您可以选择常用框架，创建训练作业来构建模型。

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。
- 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，确保帐户未欠费。
- 确保您使用的OBS目录与ModelArts在同一区域。

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。
- 训练脚本中，“数据来源”、“训练输出位置”，两个参数必须为OBS路径。当需要对OBS进行读写交互时，建议使用**MoXing接口**进行读写操作。

训练管理支持的 AI 常用框架

当前ModelArts支持的AI引擎及对应版本如下所示。

表 3-6 旧版训练作业支持的 AI 引擎

工作环境	适配芯片	系统架构	系统版本	AI引擎与版本	支持的cuda或Ascend版本
TensorFlow	CPU/GPU	x86_64	Ubuntu16.04	TF-1.8.0-python3.6	-
				TF-1.13.1-python3.6	-
				tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64	cuda10.1
MXNet	CPU/GPU	x86_64	Ubuntu16.04	MXNet-1.2.1-python3.6	-
Spark_MLlib	CPU	x86_64	Ubuntu16.04	Spark-2.3.2-python3.6	-
Ray	CPU/GPU	x86_64	Ubuntu16.04	RAY-0.7.4-python3.6	-
XGBoost-Sklearn	CPU	x86_64	Ubuntu16.04	XGBoost-0.80-Sklearn-0.18.1-python2.7	-
				XGBoost-0.80-Sklearn-0.18.1-python3.6	-
PyTorch	CPU/GPU	x86_64	Ubuntu16.04	PyTorch-1.0.0-python3.6	-
				PyTorch-1.3.0-python3.6	-
				PyTorch-1.4.0-python3.6	-

工作环境	适配芯片	系统架构	系统版本	AI引擎与版本	支持的cuda或Ascend版本
				pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64	cuda10.2
Ascend-Powered-Engine	Ascend	aarch64	Euler2.8	mindspore_1.7.0-cann_5.1.0-py_3.7-euler_2.8.3-aarch64	5.1.0
				tensorflow_1.15-cann_5.1.0-py_3.7-euler_2.8.3-aarch64	5.1.0
MPI	CPU/GPU	x86_64	Ubuntu18.04	mindspore_1.3.0-cuda_10.1-py_3.7-ubuntu_1804-x86_64	cuda10.1
Caffe	CPU/GPU	x86_64	Ubuntu16.04	Caffe-1.0.0-python2.7	cuda8.0

说明

- MoXing是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。
- **Ascend-Powered-Engine**仅在“华北-北京四”区域支持。
- 如果您的训练脚本支持的版本与训练支持的AI引擎提供的版本存在差异，会出现训练失败的情况

创建训练作业

1. 登录ModelArts管理控制台，在左侧导航栏中选择“训练管理 > 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。您可以根据实际情况填写“名称”和“描述”信息。“版本”信息由系统自动生成，按“V0001”、“V0002”规则命名，用户无法修改。
   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3-7**。

图 3-6 算法来源为常用框架

表 3-7 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。
算法来源	常用框架	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。 目前支持的AI引擎及其版本请参见**训练管理支持的AI常用框架**。 如果您的模型需要安装Python依赖包时，请按照ModelArts定义的要求将依赖包及其配置文件放置“代码目录”中，详细说明请参见**模型中引用依赖包时，如何创建训练作业?**

参数名称	子参数	说明
数据来源	数据集	从ModelArts数据管理中选择可用的数据集及其版本。 ● “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。 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。
	数据存储位置	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。 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。
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。
运行参数	-	代码中的命令行参数设置值，请根据您编写的算法代码逻辑进行填写，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。
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 3-7 选择训练作业资源

表 3-8 资源参数说明

参数名称	说明
资源池	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”。公共资源池又可以选择CPU或GPU两种类型，不同类型的资源池，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（旧版即将下线）**。
类型	当选择公共资源池时，需选择资源类型。目前支持CPU和GPU两种类型。 GPU资源性能更佳，CPU资源性价比更高。如果您选择的算法已定义使用CPU或GPU，界面将自动呈现此资源类型，请务必根据要求选择。 不同的资源类型的数据盘容量是不同的，详细介绍参考**训练环境中不同规格资源“/cache”目录的大小**。 **说明** 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。
规格	针对不同的资源类型，选择资源规格。 其中针对“已有算法”、“常用框架”或“自定义镜像”，ModelArts支持使用Ascend创建训练作业，且此资源仅在“华北-北京四”可用。 前缀带“[限时免费]”信息的为免费规格，您可以使用此规格免费体验ModelArts的训练作业功能。此规格的使用注意事项，请参见**免费体验AI全流程开发**。
计算节点个数	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。 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。

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 3-8 配置训练作业订阅消息

表 3-9 订阅消息及作业参数的参数说明

参数名称	说明
订阅消息	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。 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。 ● “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	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。 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**。

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“提交”，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。

3.4.4 使用自定义镜像训练模型

如果您开发算法时使用的框架并不是常用框架，您可以将算法构建为一个自定义镜像，通过自定义镜像创建训练作业。

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“自定义”，请按照规范完成镜像制作，并上传至SWR服务，请参考**训练作业自定义镜像规范**。
- 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，确保帐户未欠费。
- 确保您使用的OBS目录与ModelArts在同一区域。

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。
- 训练脚本中，“数据来源”、“训练输出位置”，两个参数必须为OBS路径。当需要对路径中进行读写交互时，建议使用**MoXing接口**进行读写操作。

创建训练作业

1. 登录ModelArts管理控制台，在左侧导航栏中选择“训练管理 > 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。
      您可以根据实际情况填写“名称”和“描述”信息。
   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3-10**。

表 3-10 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
算法来源	自定义	自定义镜像的相关规范请参见训练作业自定义镜像规范。 ● “镜像地址”：镜像上传到SWR后生成的地址。 ● “代码目录”：训练代码文件存储的OBS路径。 ● “运行命令”：镜像启动后的运行命令，根据实际情况填写。
数据来源	数据集	从ModelArts数据管理中选择可用的数据集及其版本。 ● “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。
	数据存储位置	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。
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。
环境变量	-	请根据您的镜像文件，添加环境变量，此参数为可选。单击“增加环境变量”可增加多个变量参数。
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。

表 3-11 资源参数说明

参数名称	说明
资源池	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”。公共资源池又可以选择CPU或GPU两种类型，不同类型的资源池，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（旧版即将下线）。

参数名称	说明
类型	当选择公共资源池时，选择资源类型。目前支持CPU和GPU两种类型。 GPU资源性能更佳，CPU资源性价比更高。如果您选择的算法已定义使用CPU或GPU，界面将自动呈现此资源类型，请务必根据要求选择。 不同的资源类型的数据盘容量是不同的，详细介绍参考**训练环境中不同规格资源“/cache”目录的大小**。 **说明** • 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。
规格	针对不同的资源类型，选择资源规格。 其中针对“已有算法”、“常用框架”或“自定义镜像”，ModelArts支持使用Ascend创建训练作业，且此资源仅在“华北-北京四”可用。 资源规格中，前缀带“[限时免费]”信息的为免费规格，您可以使用此规格免费体验ModelArts的训练作业功能。此规格的使用注意事项，请参见**免费体验AI全流程开发**。
计算节点个数	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。

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 3-9 配置训练作业订阅消息

表 3-12 订阅消息及作业参数参数说明

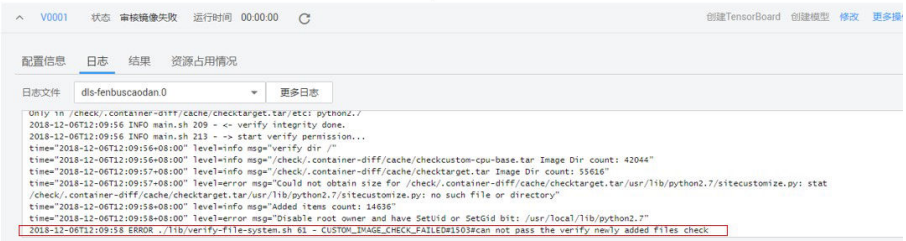
参数名称	说明
订阅消息	订阅消息使用消息通知服务，在事件列表中选择需要监控的资源池状态，在事件发生时发送消息通知。 此参数为可选参数，您可以根据实际情况设置是否打开开关。如果开启订阅消息，请根据实际情况填写如下参数。 ● “主题名”：订阅消息主题名称。您可以单击创建主题，在消息通知服务中创建主题。 ● “事件列表”：订阅事件。当前可选择“OnJobRunning”、“OnJobSucceeded”、“OnJobFailed”三种事件，分别代表训练运行中、运行成功、运行失败。
保存作业参数	勾选此参数，表示将当前训练作业设置的作业参数保存，方便后续一键复制使用。 勾选“保存训练参数”，然后填写“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**。

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“提交”，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。

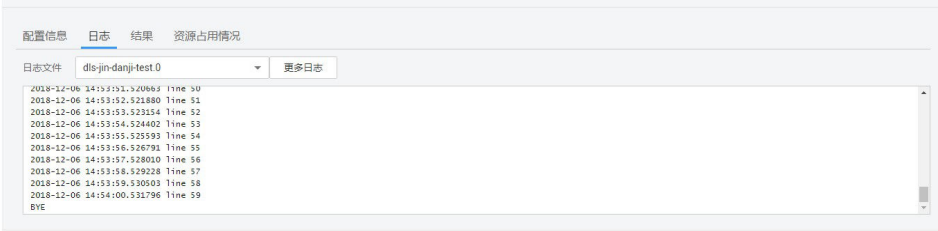
   完成自定义镜像作业创建后，系统默认授权ModelArts获取镜像运行。在第一次运行自定义镜像作业时，ModelArts进行自定义镜像审核，审核内容请参见**训练作业自定义镜像规范**，审核失败的原因可在日志查看，用户根据日志做相应的修改。

   图 3-10 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，跑自定义镜像训练作业。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。

图 3-11 运行日志



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。
- 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。
- 自定义镜像的默认用户必须为“uid”为“1101”的用户。

3.5 停止或删除作业

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。

3.6 管理训练作业版本

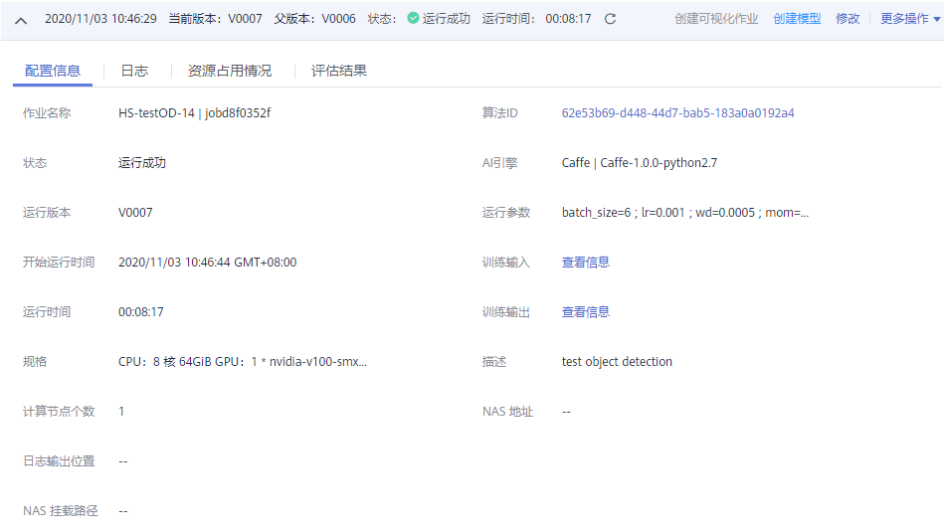
在模型构建过程中，您可能需要根据训练结果，不停的调整数据、训练参数或模型，以获得一个满意的模型。因此，ModelArts为了方便用户在调整内容后快速高效的训练模型，提供了管理训练作业版本的能力。每训练一次，生成一个版本，不同的作业版本之间，能快速进行对比，获得对比结果。

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“训练管理 > 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。单击版本左侧的小三角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 3-12 查看训练作业版本



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”。

说明

使用预置算法创建的训练作业，才会显示其对应的“F1值”、“召回率”、“精确率”、“准确率”。针对使用常用框架、或自定义镜像创建的训练作业，请在您的训练脚本代码中定义好这些参数的输出，暂不支持在界面中查看。

图 3-13 训练版本对比

训练版本对比

版本	运行参数	F1值	召回率	精确率	准确率
V0006	split_spec=train:0.8,eval:0.2 num_gpus=1 batch_size=32 eval_batch_size=32 learning_rate_strategy=0.002 evaluate_every_n_epochs=1 save_interval_secs=2000000 max_epoches=100 log_every_n_steps=10 save_summaries_steps=5	0.054877	0.998818	0.028213	0.944444
V0005	split_spec=train:0.8,eval:0.2 num_gpus=1 batch_size=32 eval_batch_size=32 learning_rate_strategy=0.002 evaluate_every_n_epochs=1 save_interval_secs=2000000 max_epoches=100 log_every_n_steps=10 save_summaries_steps=5	0.053715	0.998822	0.027599	0.944444

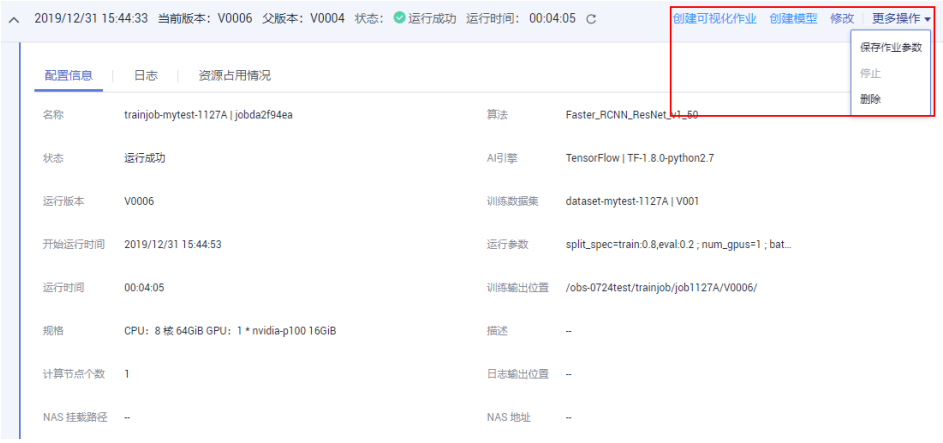
基于训练作业版本的快捷操作

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 3-13 快捷操作说明

操作	说明
创建可视化作业	基于当前训练版本创建可视化作业，详细参见**管理可视化作业**。
创建模型	基于当前训练版本创建模型，详细参见**创建AI应用**。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 3-14 快捷操作



3.7 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**和**评估结果**，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“训练管理 > 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3-14**所示。

图 3-15 训练作业详情

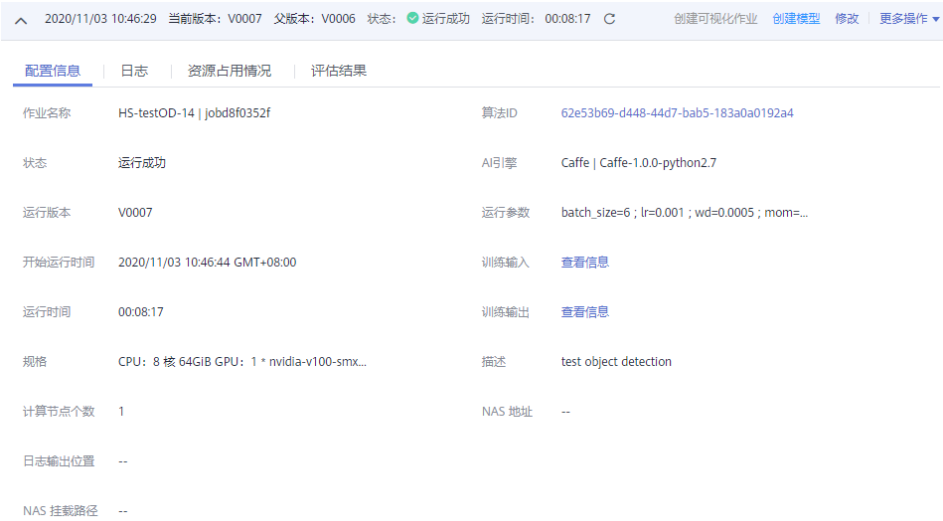


表 3-14 训练作业详情

参数	说明
当前版本	训练作业版本，由系统自动定义，命名规则为V0001、V0002。
状态	训练作业的状态。
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、NPU和内存。

评估结果

针对使用ModelArts官方发布的预置算法创建训练作业时，其训练作业详情支持查看评估结果。如果您的训练脚本中按照ModelArts规范添加了相应的评估代码，在训练作业运行结束后，也可在作业详情页面查看评估结果，添加评估代码指导请参见**添加评估结果**。

说明

当前仅图像分类-ResNet_v1_50、物体检测-FasterRCNN_ResNet50、物体检测-EfficientDet预置算法支持模型评估结果查看功能。

当训练作业运行结束，且状态为“运行成功”时，进入作业详情页面，可在“评估结果”页签下查看详情，如**图3-16**所示。

评估结果包含“评估综述”、“精度评估”、“敏感度分析”等维度，包含了常用的模型指标。由于每个模型情况不同，系统将自动根据您的模型指标情况，给出一些调优建议，请仔细阅读界面中的建议和指导，对您的模型进行进一步的调优。

图 3-16 评估结果

3.8 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“训练管理 > 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。

编辑作业参数

1. 登录ModelArts管理控制台，在左侧导航栏中选择“训练管理 > 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，参见创建训练作业章节，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“训练管理 > 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。

3.9 添加评估结果

训练作业运行结束后，ModelArts将自动为您的模型进行评估，并且给出调优诊断和建议，详细功能描述请参见评估结果。

- 针对使用预置算法创建训练作业，无需任何配置，即可查看此评估结果。
- 针对用户自己编写训练脚本或自定义镜像方式创建的训练作业，则需要在您的训练代码中添加评估代码，才可以在训练作业结束后查看相应的评估诊断建议。

说明

1. 只支持验证集的数据格式为图片
2. 目前，仅如下常用框架的训练脚本支持添加评估代码。
   - TF-1.13.1-python3.6
   - TF-2.1.0-python3.6
   - PyTorch-1.4.0-python3.6

本章节介绍如何在训练中使用评估代码。对训练代码做一定的适配和修正，分为三个方面：**添加输出目录**、**拷贝数据集到本地**、**映射数据集路径到OBS**。

添加输出目录

添加输出目录的代码比较简单，即在代码中添加一个输出评估结果文件的目录，被称为train_url，也就是页面上的训练输出位置。并把train_url添加到使用的函数analysis中，使用save_path来获取train_url。示例代码如下所示：

```
FLAGS = tf.app.flags.FLAGS
tf.app.flags.DEFINE_string('model_url', '', 'path to saved model')
tf.app.flags.DEFINE_string('data_url', '', 'path to output files')
tf.app.flags.DEFINE_string('train_url', '', 'path to output files')
tf.app.flags.DEFINE_string('adv_param_json',
                           '{"attack_method":"FGSM","eps":40}',
                           'params for adversarial attacks')
FLAGS(sys.argv, known_only=True)

...

# analyse
res = analyse(
    task_type=task_type,
    pred_list=pred_list,
    label_list=label_list,
    name_list=file_name_list,
    label_map_dict=label_dict,
    save_path=FLAGS.train_url)
```

拷贝数据集到本地

拷贝数据集到本地主要是为了防止长时间访问OBS容易导致OBS连接中断使得作业卡住，所以一般先将数据拷贝到本地再进行操作。

数据集拷贝有两种方式，推荐使用OBS路径进行处理拷贝。

- OBS路径（推荐）

  直接使用moxing的copy_parallel接口，拷贝对应的OBS路径。

- ModelArts数据管理中的数据集（即manifest文件格式）

  使用moxing的copy_manifest接口将文件拷贝到本地并获取新的manifest文件路径，然后使用SDK解析新的manifest文件。

```
if data_path.startswith('obs://'):
    if '.manifest' in data_path:
        new_manifest_path, _ = mox.file.copy_manifest(data_path, '/cache/data/')
        data_path = new_manifest_path
    else:
        mox.file.copy_parallel(data_path, '/cache/data/')
        data_path = '/cache/data/'
    print('------------- download dataset success ------------')
```

映射数据集路径到 OBS

由于最终JSON体中需要填写的是图片文件的真实路径，也就是OBS对应的路径，所以在拷贝到本地做完分析和评估操作后，需要将原本的本地数据集路径映射到OBS路径，然后将新的list送入analysis接口。

如果使用的是OBS路径作为输入的data_url，则只需要替换本地路径的字符串即可。

```
if FLAGS.data_url.startswith('obs://'):
    for idx, item in enumerate(file_name_list):
        file_name_list[idx] = item.replace(data_path, FLAGS.data_url)
```

如果使用manifest文件，需要再解析一遍原版的manifest文件获取list，然后再送入analysis接口。

```
if or FLAGS.data_url.startswith('obs://'):
    if 'manifest' in FLAGS.data_url:
            file_name_list = []
            manifest, _ = get_sample_list(
                manifest_path=FLAGS.data_url, task_type='image_classification')
            for item in manifest:
                if len(item[1]) != 0:
                    file_name_list.append(item[0])
```

完整的适配了训练作业创建的图像分类样例代码如下：

```
import json
import logging
import os
import sys
import tempfile

import h5py
import numpy as np
from PIL import Image

import moxing as mox
import tensorflow as tf
from deep_moxing.framework.manifest_api.manifest_api import get_sample_list
from deep_moxing.model_analysis.api import analyse, tmp_save
from deep_moxing.model_analysis.common.constant import TMP_FILE_NAME

logging.basicConfig(level=logging.DEBUG)

FLAGS = tf.app.flags.FLAGS
tf.app.flags.DEFINE_string('model_url', '', 'path to saved model')
tf.app.flags.DEFINE_string('data_url', '', 'path to output files')
tf.app.flags.DEFINE_string('train_url', '', 'path to output files')
tf.app.flags.DEFINE_string('adv_param_json',
                           '{"attack_method":"FGSM","eps":40}',
                           'params for adversarial attacks')
FLAGS(sys.argv, known_only=True)


def _preprocess(data_path):
    img = Image.open(data_path)
    img = img.convert('RGB')
    img = np.asarray(img, dtype=np.float32)
    img = img[np.newaxis, :, :, :]
    return img


def softmax(x):
    x = np.array(x)
    orig_shape = x.shape
    if len(x.shape) > 1:
        # Matrix
        x = np.apply_along_axis(lambda x: np.exp(x - np.max(x)), 1, x)
```

```
        denominator = np.apply_along_axis(lambda x: 1.0 / np.sum(x), 1, x)
        if len(denominator.shape) == 1:
            denominator = denominator.reshape((denominator.shape[0], 1))
        x = x * denominator
    else:
        # Vector
        x_max = np.max(x)
        x = x - x_max
        numerator = np.exp(x)
        denominator = 1.0 / np.sum(numerator)
        x = numerator.dot(denominator)
    assert x.shape == orig_shape
    return x


def get_dataset(data_path, label_map_dict):
    label_list = []
    img_name_list = []
    if 'manifest' in data_path:
        manifest, _ = get_sample_list(
            manifest_path=data_path, task_type='image_classification')
        for item in manifest:
            if len(item[1]) != 0:
                label_list.append(label_map_dict.get(item[1][0]))
                img_name_list.append(item[0])
            else:
                continue
    else:
        label_name_list = os.listdir(data_path)
        label_dict = {}
        for idx, item in enumerate(label_name_list):
            label_dict[str(idx)] = item
            sub_img_list = os.listdir(os.path.join(data_path, item))
            img_name_list += [
                os.path.join(data_path, item, img_name) for img_name in sub_img_list
            ]
            label_list += [label_map_dict.get(item)] * len(sub_img_list)
    return img_name_list, label_list


def deal_ckpt_and_data_with_obs():
    pb_dir = FLAGS.model_url
    data_path = FLAGS.data_url

    if pb_dir.startswith('obs://'):
        mox.file.copy_parallel(pb_dir, '/cache/ckpt/')
        pb_dir = '/cache/ckpt'
        print('------------- download success -------------')
    if data_path.startswith('obs://'):
        if '.manifest' in data_path:
            new_manifest_path, _ = mox.file.copy_manifest(data_path, '/cache/data/')
            data_path = new_manifest_path
        else:
            mox.file.copy_parallel(data_path, '/cache/data/')
            data_path = '/cache/data/'
        print('------------- download dataset success -------------')
    assert os.path.isdir(pb_dir), 'Error, pb_dir must be a directory'
    return pb_dir, data_path


def evalution():
    pb_dir, data_path = deal_ckpt_and_data_with_obs()
    index_file = os.path.join(pb_dir, 'index')
    try:
        label_file = h5py.File(index_file, 'r')
        label_array = label_file['labels_list'][:].tolist()
        label_array = [item.decode('utf-8') for item in label_array]
    except Exception as e:
        logging.warning(e)
```

```
        logging.warning('index file is not a h5 file, try json.')
        with open(index_file, 'r') as load_f:
            label_file = json.load(load_f)
        label_array = label_file['labels_list'][:]
    label_map_dict = {}
    label_dict = {}
    for idx, item in enumerate(label_array):
        label_map_dict[item] = idx
        label_dict[idx] = item
    print(label_map_dict)
    print(label_dict)

    data_file_list, label_list = get_dataset(data_path, label_map_dict)

    assert len(label_list) > 0, 'missing valid data'
    assert None not in label_list, 'dataset and model not match'

    pred_list = []
    file_name_list = []
    img_list = []

    for img_path in data_file_list:
        img = _preprocess(img_path)
        img_list.append(img)
        file_name_list.append(img_path)

    config = tf.ConfigProto()
    config.gpu_options.allow_growth = True
    config.gpu_options.visible_device_list = '0'
    with tf.Session(graph=tf.Graph(), config=config) as sess:
        meta_graph_def = tf.saved_model.loader.load(
            sess, [tf.saved_model.tag_constants.SERVING], pb_dir)
        signature = meta_graph_def.signature_def
        signature_key = 'predict_object'
        input_key = 'images'
        output_key = 'logits'
        x_tensor_name = signature[signature_key].inputs[input_key].name
        y_tensor_name = signature[signature_key].outputs[output_key].name
        x = sess.graph.get_tensor_by_name(x_tensor_name)
        y = sess.graph.get_tensor_by_name(y_tensor_name)
        for img in img_list:
            pred_output = sess.run([y], {x: img})
            pred_output = softmax(pred_output[0])
            pred_list.append(pred_output[0].tolist())

    label_dict = json.dumps(label_dict)
    task_type = 'image_classification'

    if FLAGS.data_url.startswith('obs://'):
        if 'manifest' in FLAGS.data_url:
            file_name_list = []
            manifest, _ = get_sample_list(
                manifest_path=FLAGS.data_url, task_type='image_classification')
            for item in manifest:
                if len(item[1]) != 0:
                    file_name_list.append(item[0])
        for idx, item in enumerate(file_name_list):
            file_name_list[idx] = item.replace(data_path, FLAGS.data_url)
    # analyse
    res = analyse(
        task_type=task_type,
        pred_list=pred_list,
        label_list=label_list,
        name_list=file_name_list,
        label_map_dict=label_dict,
        save_path=FLAGS.train_url)

if __name__ == "__main__":
    evalution()
```

3.10 管理可视化作业

当前，您管理的ModelArts可视化作业支持创建TensorBoard类型和MindInsight两种类型。

TensorBoard和MindInsight能够有效地展示训练作业在运行过程中的变化趋势以及训练中使用到的数据信息。

- 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**。

- MindInsight

  MindInsight能可视化展现出训练过程中的标量、图像、计算图以及模型超参等信息，同时提供训练看板、模型溯源、数据溯源、性能调试等功能，帮助您在更高效地训练调试模型。MindInsight当前支持基于MindSpore引擎的训练作业。MindInsight相关概念请参考**MindSpore官网**。

您可以使用模型训练时产生的Summary文件来创建可视化作业。

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MindSpore引擎编写程序时

  MindSpore支持将数据信息保存到Summary日志文件中，并通过可视化界面进行展示。将数据记录到Summary日志文件中的具体方式请参考**收集Summary数据**。

- 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('OBS路径'))
```

注意事项

- 运行中的可视化作业会一直按需计费，当您不需要使用时，建议停止可视化作业，避免产生不必要的费用。可视化作业支持自动停止功能，即在指定时间后停止可视化作业，为避免产生不必要的费用，推荐启用此功能。
- 默认使用CPU资源运行可视化作业，且不支持修改为其他资源池。
- 确保您使用的OBS目录与ModelArts在同一区域。

创建可视化作业

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“可视化作业”页签。
2. 在可视化作业列表中，单击左上方“创建”，进入“创建可视化作业”界面。
3. 其中，“计费模式”设置为“按需计费”，“作业类型”为“TensorBoard”和“MindInsight”两种类型。请根据实际情况填写可视化作业“名称”、“描述”，设置“训练输出位置”和“自动停止”参数。
   - “训练输出位置”：选择创建训练作业时的“训练输出位置”。
   - “自动停止”：设置是否开启自动停止功能。由于运行中的可视化作业会一直计费，为避免产生不必要的费用，您可以开启自动停止功能，在指定时间后自动停止可视化作业。目前支持设置为“1小时后”、“2小时后”、“4小时后”、“6小时后”、“自定义”。如果选择“自定义”的模式，可在右侧输入框中输入1~24范围内的任意整数。

图 3-17 创建可视化作业

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”，完成可视化作业的创建。

   在可视化作业列表中，当状态变为“运行中”时，表示可视化作业已创建完成。您可以单击可视化作业名称进入查看详情。

打开可视化作业

在可视化作业列表中，单击可视化作业名称，即可打开可视化显示界面。只有“运行中”状态的可视化作业支持打开。

图 3-18 TensorBoard 界面

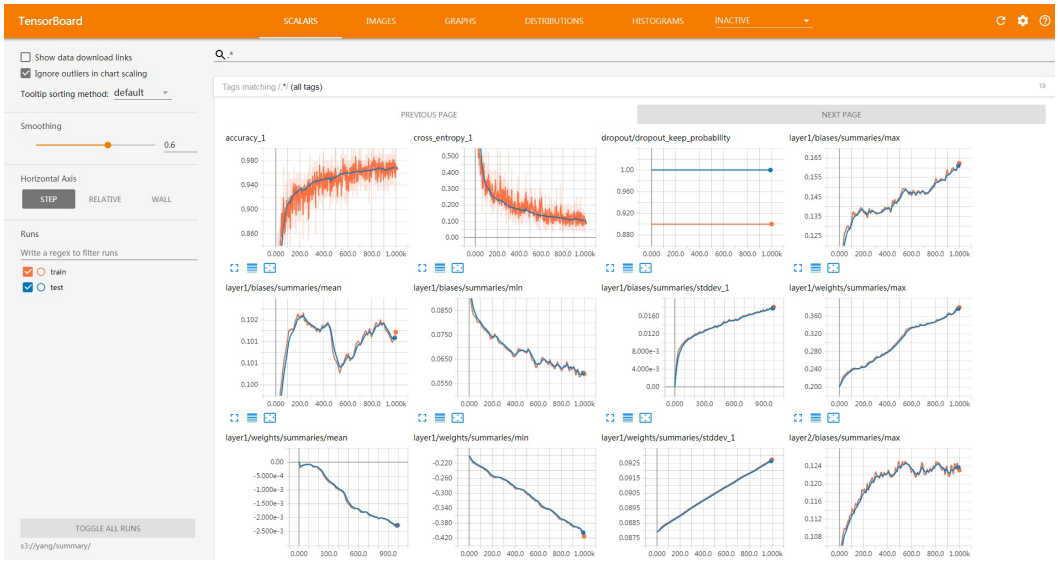
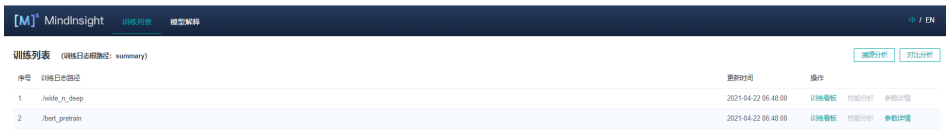


图 3-19 MindInsight 界面



运行或停止可视化作业

- **停止可视化作业**：由于“运行中”的可视化作业将一直按需计费，在不需要使用时，您可以停止可视化作业停止计费。在可视化作业列表中，单击“操作”列的“停止”，即可停止可视化作业。
- **运行可视化作业**：对于“已取消”状态的可视化作业，您可以重新运行并使用可视化作业。在可视化作业列表中，单击“操作”列的“运行”，即可运行可视化作业。

删除可视化作业

如果您的可视化作业不再使用，您可以删除可视化作业释放资源。在可视化作业列表中，单击“操作”列的“删除”，即可删除可视化作业。

说明

可视化作业删除后不可恢复，需重新创建可视化作业。请谨慎操作。

4 资源池（旧版即将下线）

ModelArts 资源池说明

说明

ModelArts已上线新版专属资源池，推荐用户使用新版专属资源池，新版专属资源池管理文档请参考资源池管理。

当前章节仅介绍了旧版专属资源池，旧版专属资源池即将下线。新旧版专属资源池的差异请参考资源池介绍。

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。

- **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。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

  创建专属资源池

  扩缩容专属资源池

  删除专属资源池

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。
- 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AI应用的部署上线。
- 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。

创建专属资源池

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”，进入创建专属资源池界面。
4. 在“创建专属资源池”界面填写参数，参数填写请参见表4-1和表4-2。

表 4-1 “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。 名称由大小写字母、数字、中划线和下划线组成。
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	请根据界面提示选择需要使用的规格，GPU性能更好，CPU更加实惠。如果某一规格出现“售罄”字样，表示此规格已用完，需要等待其他用户删除资源池后，您才可以重新购买。
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。 ● 自动续费。开通自动续费后，系统将在产品到期前自动续费，无需用户手动操作。

表 4-2 “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	选择计费模式，“包年/包月”或“按需计费”。“包年/包月”仅在北京四支持。
名称	专属资源池的名称。 名称只能以小写字母开头，由小写字母、数字、中划线组成，不能以中划线结尾，长度为4~24个字符。
描述	专属资源池的简要描述。

参数名称	说明
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。
可用区	您可以根据实际情况选择“随机分配”、“可用区1”、“可用区2”、“可用区3”。可用区是在同一区域下，电力、网络隔离的物理区域。可用区之间内网互通，不同可用区之间物理隔离。如果您需要提高工作负载的高可靠性，建议您将云服务器创建在不同的可用区。
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	请根据界面提示选择需要使用的规格，GPU性能更好，CPU更加实惠。如果某一规格出现“售罄”字样，表示此规格已用完，需要等待其他用户删除资源池后，您才可以重新购买。
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11个月。 • 自动续费。开通自动续费后，系统将在产品到期前自动续费，无需用户手动操作。

5. 规格确认无误后，根据界面提示完成专属资源池的创建。当专属资源池创建成功后，其状态将变为“运行中”。

须知

专属资源池购买后，需要初始化环境后才能使用。

（可选）打通 VPC

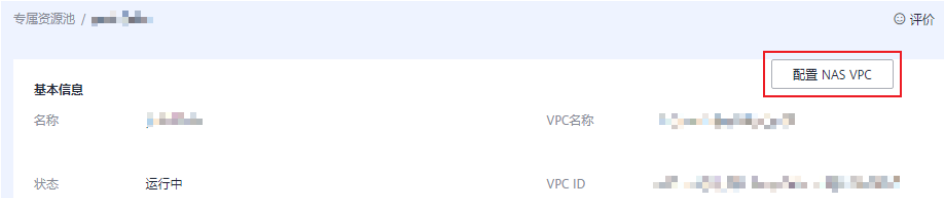
资源池创建成功后，可在资源池详情页打通VPC。具体步骤如下：

说明

仅“开发环境/训练专用”的专属资源池支持打通VPC。

1. 进入专属资源池管理页面，在专属资源池所在行，单击资源池名称，进入专属资源池详情页。
2. 在专属资源池详情页，单击“配置NAS VPC”。

图 4-1 配置 NAS VPC



3. 在修改NAS VPC页面，打开“连通NAS VPC”，设置NAS VPC和NAS子网。

图 4-2 修改 NAS VPC

   - 如果没有VPC可选，可以单击右侧的“创建虚拟私有云”，跳转到网络控制台，申请创建虚拟私有云。
   - 如果没有子网可选，可以单击右侧的“创建子网”，跳转到网络控制台，创建可用的子网。
4. 设置完成后，即可打通VPC。

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。

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，请务必选择当前帐号的配额，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4-3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。

注意

部署上线类资源池用于运行推理部署在线服务，缩容节点时请务必保证该节点上无运行实例，否则可能影响到您的线上业务。如果不确定待缩容节点上是否有实例在运行，请提咨询单咨询。

图 4-3 缩容时选择删除节点

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。

说明

缩容节点时，提交缩容操作后，删除节点操作不是立即成功的，还需要后台做处理，此时请勿去专属资源池列表中删除节点，可能会导致删除失败。

提交缩容操作后，可以在专属资源池的详情页中查看事件列表。Begin to delete resource node %s表示开始删除节点；当出现事件Resource node %s deleted，表示缩容节点在后台删除成功。

专属资源池计费模式转换

- “按需计费”的专属资源池转换为“包年包月”的专属资源池

  ModelArts支持训练的专属资源池从“按需计费”转换为“包年包月”。前提是您需要使用按需计费的专属资源池完成一次以上的消费。如果购买的按需计费专属资源池没有消费记录，转周期操作会报错。

  转包周期的操作步骤如下所示：

  a. 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“转包周期”，进入“按需转包年按月”页面。

  b. 根据页面提示选择包年包月的时长并勾选是否自动续费，完成支付后即可转换成功。

- “包年包月”的专属资源池转换为“按需计费”的专属资源池

  针对用于训练的“包年包月”专属资源池，支持在包周期到期后，将专属资源池设置为按需计费。

  转按需的操作步骤如下：

  a. 进入ModelArts控制台，单击上方设置栏的“费用中心>续费管理”，进入续费管理页面。

  图 4-4 进入费用中心

  b. 进入续费管理页面，在专属资源池所在行，单击操作列“更多>到期转按需”。

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。

说明

- 专属资源池删除后，将导致使用此资源的训练作业、Notebook、在线服务和批量服务等不可用，且删除后不可恢复，请谨慎操作。

1. 进入专属资源池管理页面，释放资源。
   - 如果是“包年/包月”类型的专属资源池，请单击操作列“退订”完成订单退费，退订完成后将自动删除该资源。
   - 如果是“按需”计费类型的专属资源池，单击操作列的“删除”即可。
   - 对于创建失败的专属资源池，请单击操作列“删除”完成资源删除。
2. 在弹出的确认对话框中，单击“确定”，完成资源删除。

5 使用自定义镜像

5.1 自定义镜像简介

ModelArts为用户提供了多种常见的预置引擎，但是当用户对深度学习引擎、开发库有特殊需求场景的时候，预置AI引擎已经不能满足用户需求。ModelArts提供自定义镜像功能支持用户自定义运行引擎。

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，灵活性比较高，便于支持任意计算引擎的作业启动需求。

关联服务介绍

使用自定义镜像功能可能涉及以下云服务：容器镜像服务、对象存储服务、弹性云服务器。

- 容器镜像服务

  容器镜像服务（Software Repository for Container，SWR）是一种支持镜像全生命周期管理的服务， 提供简单易用、安全可靠的镜像管理功能，帮助您快速部署容器化服务。您可以通过界面、社区CLI和原生API上传、下载和管理容器镜像。

  ModelArts训练和创建AI应用使用的自定义镜像需要从SWR服务管理列表获取。您制作的自定义镜像需要上传至SWR服务。

  图 5-1 获取镜像列表

- 对象存储服务

  对象存储服务（Object Storage Service，OBS）是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力。

  在创建训练作业和创建AI应用时往往存在数据交互，您需要的数据可以存储至OBS服务。

- 弹性云服务器

  弹性云服务器（Elastic Cloud Server，ECS）是由CPU、内存、操作系统、云硬盘组成的基础的计算组件。弹性云服务器创建成功后，您就可以像使用自己的本地PC或物理服务器一样，在云上使用弹性云服务器。

  在制作自定义镜像时，您可以在本地环境或者ECS上完成自定义镜像制作。

说明

在您使用自定义镜像功能时，ModelArts可能需要访问您的容器镜像服务SWR、对象存储服务OBS等依赖服务，若没有授权，这些功能将不能正常使用。建议您使用委托授权功能，将依赖服务操作权限委托给ModelArts服务，让ModelArts以您的身份使用依赖服务，代替您进行一些资源操作。详细操作参见使用委托授权。

自定义镜像使用场景

- **用于训练模型**

  如果您已经在本地完成模型开发或训练脚本的开发，且您使用的AI引擎是ModelArts不支持的框架。您可以基于ModelArts提供的基础镜像包制作自定义镜像，并上传至SWR服务。您可以在ModelArts使用此自定义镜像创建训练作业，使用ModelArts提供的资源训练模型。

- **用于创建AI应用**

  如果您使用了ModelArts不支持的AI引擎开发模型，也可通过制作自定义镜像，导入ModelArts创建为AI应用，并支持进行统一管理和部署为服务。

自定义镜像的制作流程

1. 购买弹性云服务器或者应用本地主机搭建Docker环境。
2. 在本地环境拉取基础镜像。
3. 根据您的实际需求编写Dockerfile文件构建自定义镜像。如何高效编写Dockerfile指导可参考**SWR服务最佳实践**。
   - 如果您使用自定义镜像用于训练作业请参考示例**训练作业自定义镜像规范**。
   - 如果您使用自定义镜像用于创建AI应用请参考示例**创建AI应用的自定义镜像规范**。
4. 当完成自定义镜像制作后，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。

5.2 制作和上传自定义镜像

ModelArts支持您使用自定义镜像创建训练作业和创建AI应用。在制作和上传自定义镜像之前，您需要了解以下内容：

- SWR服务。

  容器镜像服务（Software Repository for Container，简称SWR）是一种支持镜像全生命周期管理的服务， 提供简单易用、安全可靠的镜像管理功能，帮助您快速部署容器化服务。您可以通过界面、社区CLI和原生API上传、下载和管理容器镜像。

  ModelArts训练和创建AI应用使用的自定义镜像需要从SWR服务管理列表获取。您制作的自定义镜像需要上传至SWR服务。

- 自定义镜像规范。如果您使用自定义镜像用于训练作业请参考**训练作业自定义镜像规范**；如果您使用自定义镜像用于创建AI应用请参考**创建AI应用的自定义镜像规范**。

制作并上传自定义镜像

1. 购买华为云ECS或者应用本地主机搭建Docker环境。
2. 在本地环境拉取基础镜像。
3. 根据您的实际需求编写Dockerfile文件构建自定义镜像。如何高效编写Dockerfile指导可参考**SWR服务最佳实践**。
   - 如果您使用自定义镜像用于训练作业请参考示例**制作自定义镜像**。
4. 当完成自定义镜像制作后，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。

5.3 用于训练模型（旧版即将下线）

5.3.1 训练作业自定义镜像规范

说明

本章节介绍的是基于旧版训练的自定义镜像训练模型，旧版训练仅对部分存量用户可见，新用户不可见，新用户推荐使用**新版训练功能**。

针对您本地开发的模型及训练脚本，在制作镜像时，需满足ModelArts定义的规范。

规范要求

- 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。
- 不可以新增属主为“root”且权限包含“setuid”或“setgid”位的文件。
- 自定义镜像大小不能超过9.5GB。
- 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。
- 自定义镜像的默认用户必须为“uid”为“1101”的用户。
- 自定义镜像可以基于ModelArts官方提供的基础镜像制作，基础镜像请参考**基础镜像包概述**。

基础镜像包概述

为了方便用户实现代码下载、训练日志输出、上传日志文件至OBS等功能，ModelArts提供基础镜像包用于自定义镜像的制作。ModelArts提供的基础镜像有以下特点：

- 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。
- ModelArts将持续更新基础镜像版本，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，并且审核日志中出现基础镜像不匹配的错误信息时，需要使用新的基础镜像重新制作镜像。

您可以通过以下命令获取ModelArts镜像：

```
docker pull <基础镜像地址>
```

完成自定义镜像制作后，执行以下命令推送镜像至SWR服务。前提条件是已经**创建组织**并**获取SWR登录指令**。

```
docker push  swr.<region>.myhuaweicloud.com/<用户镜像所属组织>/<镜像名称>
```

根据芯片不同的需求，您可以根据实际需求获取不同的基础镜像：

- **适用于CPU的基础镜像**
- **适用于GPU的基础镜像**
- **适用于Ascend芯片基础镜像**

适用于 CPU 的基础镜像

基础镜像地址

```
swr.<region>.myhuaweicloud.com/modelarts-job-dev-image/custom-cpu-base:1.3
```

表 5-1 可选参数范围

参数	可选值	说明
<region>	• cn-north-1 • cn-north-4	镜像所在的区域。支持的值中，分别表示： • 北京一 • 北京四

基础镜像包含**表5-2**和**表5-3**

表 5-2 组件列表

名称	说明
run_train.sh	训练启动引导脚本。实现了代码目录下载，执行训练命令、重定向训练日志输出、以及训练命令结束后上传日志文件至OBS的功能。

表 5-3 工具列表

工具名称	说明
utils.sh	工具脚本。“run_train.sh”脚本依赖此脚本。 提供了SK解密，代码目录下载，日志文件上传等方法。

工具名称	说明
ip_mapper.py	网卡地址获取脚本。 默认获取ib0网卡地址IP，训练代码可以使用ib0网卡的IP加速网络通信。
dls-downloader.py	OBS下载脚本。“utils.sh”脚本依赖此脚本。

适用于 GPU 的基础镜像

- cuda10.0/10.1/10.2版本的镜像，以ubuntu18.04为基础镜像默认预装moxing
  ```
  swr.<region>.myhuaweicloud.com/modelarts-job-dev-image/custom-base-<cuda version>-<python version>-<os>-<arch>:<image tag>
  ```
- cuda8/9/92版本的镜像，默认预装Moxing
  ```
  swr.<region>.myhuaweicloud.com/modelarts-job-dev-image/custom-gpu-<cuda version>-inner-moxing-<python version>:<image tag>
  ```
- cuda8/9/92版本的镜像
  ```
  swr.<region>.myhuaweicloud.com/modelarts-job-dev-image/custom-gpu-<cuda version>-base:<image tag>
  ```

表 5-4 名称可选参数范围

参数	支持的值	说明
<region>	• cn-north-1 • cn-north-4	镜像所在的区域。支持的值中，分别表示： • 北京一 • 北京四
<cuda version>	• cuda92 • cuda9 • cuda8 • cuda10.0 • cuda10.1 • cuda10.2	镜像中已安装的CUDA版本。 **说明** 请您确认对应的CUDA版本，版本指定之后不支持更换， 否则会导致训练失败。
<image tag>	• 1.1 • 1.3	镜像版本。 • cuda8/9/92版本的镜像，支持1.3。 • cuda10.0/10.1/10.2版本的镜像支持1.1。
python version	• cp27 • cp36	python 环境。
os	ubuntu18.04	操作系统。
arch	x86	架构。

例如，在“华北-北京一”区域，ModelArts支持的基础镜像列表如下，您可根据个人需求选择相应的镜像。

cuda10之前的版本：

- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-cpu-base:1.3
- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-gpu-cuda92-base:1.3
- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-gpu-cuda9-base:1.3
- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-gpu-cuda8-base:1.3
- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-gpu-cuda9-inner-moxing-cp36:1.3
- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-gpu-cuda8-inner-moxing-cp27:1.3
- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-gpu-cuda9-inner-moxing-cp27:1.3
- ...

cuda10之后的版本：

- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-base-cuda10.0-cp36-ubuntu18.04-x86:1.1
- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-base-cuda10.1-cp36-ubuntu18.04-x86:1.1
- swr.cn-north-1.myhuaweicloud.com/modelarts-job-dev-image/custom-base-cuda10.2-cp36-ubuntu18.04-x86:1.1

基础镜像包含表5-2和表5-3

表 5-5 组件列表

名称	说明
run_train.sh	训练启动引导脚本。实现了代码目录下载，执行训练命令、重定向训练日志输出、以及训练命令结束后上传日志文件至OBS的功能。

表 5-6 工具列表

工具名称	说明
utils.sh	工具脚本。“run_train.sh”脚本依赖此脚本。 提供了SK解密，代码目录下载，日志文件上传等方法。

工具名称	说明
ip_mapper.py	网卡地址获取脚本。 默认获取ib0网卡地址IP，训练代码可以使用ib0网卡的IP加速网络通信。
dls-downloader.py	OBS下载脚本。“utils.sh”脚本依赖此脚本。

适用于 Ascend 芯片基础镜像

基础镜像包含的组件、工具如表5-7和表5-8所示

表 5-7 组件列表

名称	说明
run_train.sh	可自动完成OBS代码路径下载至本地、Ascend HCCL RANK_TABLE_FILE v0.1 格式转 v1.0 格式、多卡训练进程拉起功能。

表 5-8 工具列表

工具名称	说明
utils.sh	工具脚本。“run_train.sh”脚本依赖此脚本。 提供了SK解密，代码目录下载，日志文件上传等方法。
modelarts-downloader.py	OBS下载脚本。“utils.sh”脚本依赖此脚本。

5.3.2 使用自定义镜像创建训练作业（GPU）

说明

基于自定义镜像训练模型仅适用于旧版训练模块（仅对部分存量用户可见，新用户不可见），新版训练请参见使用自定义镜像训练模型（新版训练）。

用户将自定义镜像制作完成并上传至SWR后，可在ModelArts管理控制台，使用自定义镜像创建训练作业，完成模型训练操作。

前提条件

- 已按照ModelArts规范制作自定义镜像包，使用自定义镜像创建训练作业需遵守的规范请参见训练作业自定义镜像规范。
- 已将自定义镜像上传至SWR服务，详细操作指导请参见制作并上传自定义镜像。

创建训练作业

进入ModelArts管理控制台，创建训练作业。在使用自定义镜像创建作业时，需关注“算法来源”、“环境变量”和“资源池”参数的设置。

- **“算法来源”**

  选择“自定义”页签。

  – “镜像地址”：镜像上传到SWR后生成的地址。

  图 5-2 SWR 镜像地址

  – “代码目录”：训练代码文件存储的OBS路径。

  – “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}

  “run_train.sh”为训练启动引导脚本。执行该脚本后，ModelArts将“代码目录”下的所有内容递归下载到容器本地路径，下载后的容器本地路径为“/home/work/user-job-dir/${“代码目录”的最后一级名称}/”。

  例如，训练代码文件的OBS路径为“obs://obs-bucket/new/train.py”，“代码目录”选择“obs://obs-bucket/new/”时，则下载后的容器本地代码目录对应为“/home/work/user-job-dir/new/”。下载后的容器本地训练代码路径为“/home/work/user-job-dir/new/train.py”，运行命令可以设置为：**bash /home/work/run_train.sh python /home/work/user-job-dir/new/train.py {python_file_parameter}**

  说明

  使用自定义镜像创建训练作业过程中，ModelArts 允许用户完全自定义“运行命令”。“运行命令”中提到以下两种基本格式：

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

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

  其中，“run_train.sh，”为训练启动引导脚本。用户在制作自定义镜像过程中，可以自主实现训练启动引导脚本，也可以提前将训练代码置于自定义镜像环境中，无需遵循上述两种格式，实现完全的自定义“运行命令”。

- **“环境变量”**

  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5-9所示。用户可以根据需求来确认在自己训练脚本的python中是否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 5-9 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。
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。
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”。 HOSTS环境变量的格式为“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。当资源池为8GPU规格的专属资源池时， 容器的网络类型为主机网络，并且可以使用主机IB网络加速通信。当使用其他资源池时为容器网络。 **说明** 使用主机IB网络加速通信时，IPoIB特性需要ip_mapper.py工具获取ib0网卡IP地址。

- **“资源池”**

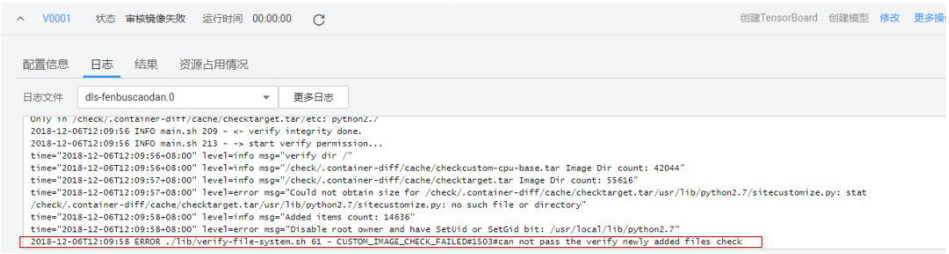
  当用户选择GPU类型的资源池时，ModelArts会挂载高速固态硬盘（NVME SSD）至“/cache”目录，用户可以使用此目录来储存临时文件。

图 5-3 创建训练作业

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**训练作业自定义镜像规范**，审核失败的原因见于日志，用户根据日志做相应的修改。

图 5-4 审核镜像失败

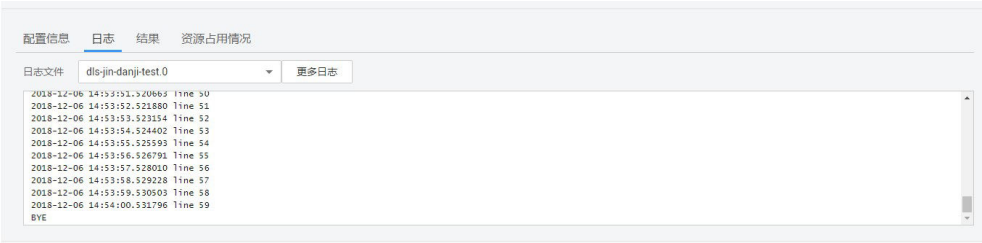


镜像审核成功后，后台就会开始启动用户自定义镜像容器，开始跑自定义镜像训练作业，用户可根据日志来查看训练情况。

说明

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。

图 5-5 运行日志



5.3.3 使用自定义镜像训练模型（Ascend）

如果 Ascend-Powered-Engine 常用框架无法满足您的需求，您可以将算法构建为一个自定义镜像，通过自定义镜像创建训练作业。

说明

基于自定义镜像训练模型仅适用于旧版训练模块（仅对部分存量用户可见，新用户不可见），新版训练请参见**使用自定义镜像训练模型（新版训练）**。

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“自定义”，请完成镜像制作，自定义镜像制作规范参见**训练作业自定义镜像规范**。
- 制作完成的自定义镜像需上传至SWR服务，请参考“《容器镜像服务用户指南》>镜像管理>客户端上传镜像”。

- 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。
- 确保您使用的OBS目录与ModelArts在同一区域。

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。
- 训练脚本中，“数据来源”、“训练输出位置”，两个参数必须为OBS路径。当需要对路径中进行读写交互时，建议使用**MoXing接口**进行读写操作。

创建训练作业

进入ModelArts管理控制台，参考**创建训练作业**操作指导，创建训练作业。在使用自定义镜像创建作业时，需关注“算法来源”、“环境变量”和“资源池”参数的设置。

您需要特别关注以下作业参数的设置：

- “镜像地址”

  镜像上传到SWR后生成的地址。

  图 5-6 SWR 镜像地址

- “运行命令”

  若使用 ModelArts Ascend 基础镜像，运行命令参考如下：

  ```
  /bin/bash /home/work/run_train.sh ${obs-code-path} ${the-base-name-of-obs-code-path}/${boot-file} '/tmp/log/train.log' ${python_file_parameter}
  ```

  run_train.sh 为ModelArts提供的启动脚本，可自动完成OBS代码路径下载至本地、Ascend HCCL RANK_TABLE_FILE v0.1 格式转 v1.0 格式、多P训练进程拉起功能，详细描述请参见**适用于Ascend芯片基础镜像**。

说明

- “obs-code-path”：OBS 代码路径，例如 obs://training-bucket/ascend-tf-1.15/resnet50/
- “the-base-name-of-obs-code-path”：OBS 代码路径的最后一级目录，例如 resnet50
- “boot-file”：以 .py 结尾的训练启动文件，例如 train.py
- '/tmp/log/train.log': 默认值，日志重定向至该文件
- “python_file_parameter”：传入训练启动文件的参数，例如 --param1=value1 --param2=value2

运行命令示例：

```
/bin/bash /home/work/run_train.sh 'obs://training-bucket/ascend-tf-1.15/resnet50/' 'resnet50/train.py' '/tmp/log/train.log' --'data_url'='obs://training-bucket/cifar-10/' --'train_url'='obs://training-bucket/model/'
```

- **“环境变量”**

表 5-10 必选环境变量说明

环境变量	说明
RANK_TABLE_FILE	该参数可以指定“jobstart_hccl.json” 文件的生成路径。 建议配置为 /user/config，则“jobstart_hccl.json ”文件路径为“/user/config/jobstart_hccl.json”。 算法开发者自行开发启动脚本时，可通过 “${RANK_TABLE_FILE}/jobstart_hccl.json”，获取文件。“jobstart_hccl.json”是 v0.1 版本的，用于分布式通信，会在运行过程中被Ascend芯片的集合通信库解析。

说明

若未添加上述环境变量，则系统不会生成 RANK TABLE FILE，训练作业日志会停留在

```
Wait for Rank table file ready
```

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5-11所示。用户可以根据需求来确认在自己训练脚本的python中是否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 5-11 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。
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。

环境变量	说明
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。
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”。 HOSTS环境变量的格式为“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。

– jobstart_hccl.json 文件格式（v0.1）示例

```
{
    "group_count": "1",
    "group_list": [{
        "device_count": "1",
        "group_name": "job-trainjob",
        "instance_count": "1",
        "instance_list": [{
            "devices": [{
                "device_id": "4",
                "device_ip": "192.1.10.254"
            }],
            "pod_name": "jobxxxxxxxx-job-trainjob-0",
            "server_id": "192.168.0.25"
        }]
    }],
    "status": "completed"
}
```

– jobstart_hccl.json 文件格式（v1.0）示例

```
{
    "server_count": "1",
    "server_list": [{
        "device": [{
            "device_id": "4",
            "device_ip": "192.1.10.254",
            "rank_id": "0"
        }],
        "server_id": "192.168.0.25"
    }],
    "status": "completed",
    "version": "1.0"
}
```

5.3.4 示例：使用自定义镜像创建训练作业

说明

本章节仅适用于指导在旧版训练模块中使用自定义镜像功能（旧版训练模块仅对部分存量用户可见，新用户不可见）。新版训练模块中使用自定义镜像功能请见**训练管理中使用自定义镜像介绍**。

本示例所需的文件存储在**Github仓库**中。本示例使用MNIST数据集，从**MNIST官网**下载。

- “mnist_softmax.py”为单机脚本。

制作并上传自定义镜像

本示例使用Dockerfile文件定制自定义镜像。

以linux x86_x64架构的主机为例，您可以购买相同规格的ECS或者应用本地已有的主机进行自定义镜像的制作。

1. 安装Docker，可参考**Docker官方文档**。

   以linux x86_64架构的操作系统为例，获取Docker安装包。使用以下指令安装Docker：

   ```
   curl -fsSL get.docker.com -o get-docker.sh
   sh get-docker.sh
   ```

   如果**docker images**命令可以执行成功，表示Docker已安装，该步骤可跳过。

2. 获取自定义镜像的基础镜像。

   训练作业的自定义镜像需要以基础镜像为基础。基础镜像名称格式参见**基础镜像包概述**。使用以下指令获取自定义镜像的基础镜像：

   ```
   docker pull swr.<region>.myhuaweicloud.com/<image org>/<image name>
   ```

   另外，您还可以使用**docker images**命令可查看本地的镜像列表。

3. 编写构建自定义镜像的Dockerfile文件。

   本示例构建tensorflow 1.13.2版本镜像。文件命名为“tf-1.13.2.dockerfile”。执行*vi tf-1.13.2.dockerfile*命令，进入文件中。

   Dockerfile文件编写的更多指导内容参见**官方指导说明**。

   ```
   FROM swr.cn-north-4.myhuaweicloud.com/modelarts-job-dev-image/custom-base-cuda10.0-cp36-ubuntu18.04-x86:1.1
   # 配置华为云的源，安装tensorflow
   RUN cp -a /etc/apt/sources.list /etc/apt/sources.list.bak && \
   sed -i "s@http://.*archive.ubuntu.com@http://repo.myhuaweicloud.com@g"  /etc/apt/sources.list && \
   sed -i "s@http://.*security.ubuntu.com@http://repo.myhuaweicloud.com@g"  /etc/apt/sources.list && \
   pip install  --trusted-host https://repo.huaweicloud.com -i https://repo.huaweicloud.com/repository/pypi/simple  tensorflow==1.13.2
   # 配置环境变量
   ENV PATH=/root/miniconda3/bin/:$PATH
   ```

4. 构建自定义镜像。

   下列例子中镜像所在的区域为cn-north-4，镜像所属组织为deep-learning-diy，在“tf-1.13.2.dockerfile”文件所在的目录执行：

   ```
   docker build -f tf-1.13.2.dockerfile . -t swr.cn-north-4.myhuaweicloud.com/deep-learning-diy/tf-1.13.2:latest
   ```

5. 推送镜像至SWR，上传镜像的详细操作可参考**SWR用户指南**。

   前提条件是已经**创建组织**并**获取SWR登录指令**。下列例子中镜像所在的区域为cn-north-4，镜像所属组织为deep-learning-diy，执行以下命令推送镜像至SWR。

   ```
   docker push swr.cn-north-4.myhuaweicloud.com/deep-learning-diy/tf-1.13.2:latest
   ```

   “swr.cn-north-4.myhuaweicloud.com/deep-learning-diy/tf-1.13.2:latest”即为此自定义镜像的“SWR_URL”。

单机训练

1. 将训练代码“mnist_softmax.py”和训练数据上传至OBS。将代码和数据都放在同一代码根目录下，以便直接下载到容器中。

   以根目录“obs://deep-learning/new/mnist/”为例：

   训练代码文件为“obs://deep-learning/new/mnist/mnist_softmax.py”；

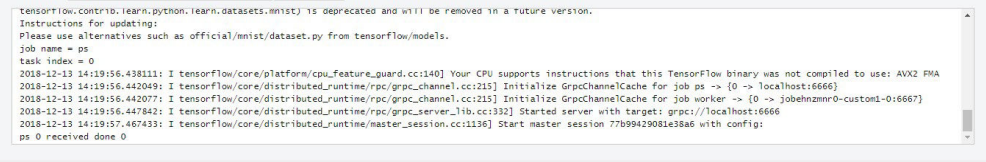
   数据存储路径为“obs://deep-learning/new/mnist/mnist_data”。

2. 创建自定义镜像训练作业，“镜像地址”、“代码目录”和“运行命令”参考如下信息填写，“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”：填写已上传镜像的“SWR_URL”。
   - “代码目录”：训练代码存储的OBS路径，即为步骤1中的代码根目录。

     在训练作业实际启动之前，ModelArts自动将“代码目录”下的所有内容递归下载到容器本地路径。下载后的容器本地路径为“/home/work/user-job-dir/${代码根目录的最后一级名称}/”。例如“代码目录”选择“obs://deep-learning/new/mnist”时，下载后的本地路径为“/home/work/user-job-dir/mnist/”，代码启动文件为“/home/work/user-job-dir/mnist/mnist_softmax.py”。
   - “运行命令”：**bash /home/work/run_train.sh python /home/work/user-job-dir/mnist/mnist_softmax.py --data_url /home/work/user-job-dir/mnist/mnist_data**

     其中，“/home/work/user-job-dir/mnist/mnist_softmax.py”为代码启动文件，“--data_url /home/work/user-job-dir/mnist/mnist_data”为数据存储路径。
3. 训练作业创建完成后，后台完成代码目录下载、自定义镜像审核以及自定义镜像的训练作业。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。程序执行成功后，日志信息如下所示。

图 5-7 运行日志信息



```
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 -> jobehnzmnr0-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 77b99429081e38a6 with config:
ps 0 received done 0
```

6 权限管理

6.1 创建并授权使用 ModelArts

如果您需要对您所拥有的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，在您的华为云帐号中，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。
- 将ModelArts资源委托给更专业、高效的其他华为云帐号或者云服务，这些帐号或者云服务可以根据权限进行代运维。

如果华为云帐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。

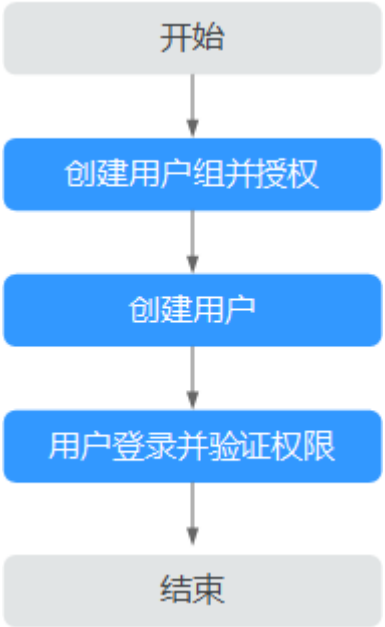
本章节为您介绍对用户授权的方法，操作流程如**图6-1**所示。

前提条件

- 给用户组授权之前，请您了解用户组可以添加的ModelArts权限，并结合实际需求进行选择，ModelArts支持的系统权限，请参见：**ModelArts系统权限**。
- 由于ModelArts的使用权限依赖OBS服务的授权，您需要为用户授予OBS的系统权限，详细说明请参见**OBS权限管理**。
- 若您需要对除ModelArts和OBS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**。

示例流程

图 6-1 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，并授予“ModelArts CommonOperations”权限。

   由于ModelArts依赖OBS权限，请为用户组授予“作用范围”为“全局级服务”的“Tenant Administrator”策略。

2. **创建用户并加入用户组**

   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，切换至授权区域，验证权限：

   - 在“服务列表”中选择ModelArts，进入ModelArts主界面，单击“专属资源池>创建”，如果无法进行创建（假设当前权限仅包含ModelArts CommonOperations），表示“ModelArts CommonOperations”已生效。
   - 在“服务列表”中选择除ModelArts外（假设当前策略仅包含ModelArts CommonOperations）的任一服务，若提示权限不足，表示“ModelArts CommonOperations”已生效。
   - 在“服务列表”中选择ModelArts，进入ModelArts主界面，单击“数据管理>数据集>创建数据集”，如果可以成功访问对应的OBS路径，表示全局级服务的“Tenant Administrator”已生效。

6.2 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**。

目前华为云支持以下两种方式创建自定义策略：

- 可视化视图创建自定义策略：无需了解策略语法，按可视化视图导航栏选择云服务、操作、资源、条件等策略内容，可自动生成策略。
- JSON视图创建自定义策略：可以在选择策略模板后，根据具体需求编辑策略内容；也可以直接在编辑框内编写JSON格式的策略内容。

具体创建步骤请参见：**创建自定义策略**。本章为您介绍常用的**ModelArts依赖的OBS权限自定义策略样例**和**ModelArts自定义策略样例**。

注意事项

- 由于ModelArts的使用权限依赖OBS服务的授权，您需要为用户授予OBS的系统权限。
- 如果一个自定义策略中包含多个服务的授权语句，这些服务必须是同一属性，即都是“全局级服务”或者“项目级服务”。
- 如果需要同时设置全局级服务和项目级服务的自定义策略，请创建两条自定义策略，“作用范围”别为“全局级服务”以及“项目级服务”，然后给用户同时授予这两条自定义策略。
- 针对当前帐号下创建的IAM用户，默认具备ModelArts的所有操作权限，当需要对IAM用户进行权限控制，如拒绝某用户使用某功能的权限时，可通过创建ModelArts自定义策略实现。

ModelArts 依赖的 OBS 权限自定义策略样例

由于ModelArts为“项目级服务”，OBS为“全局级服务”，因此需要分别创建自定义策略然后授予用户。如下示例为ModelArts依赖OBS服务的最小化权限项，包含OBS桶和OBS对象的权限。授予示例中的权限您可以通过ModelArts正常访问OBS不受限制。

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "obs:object:PutObjectAcl",
                "obs:bucket:PutBucketAcl",
                "obs:bucket:PutBucketPolicy",
                "obs:bucket:HeadBucket",
                "obs:bucket:ListAllMyBuckets",
                "obs:bucket:ListBucket",
                "obs:object:DeleteObjectVersion",
                "obs:object:AbortMultipartUpload",
                "obs:object:DeleteObject",
                "obs:object:PutObject",
                "obs:bucket:CreateBucket",
                "obs:object:GetObject",
                "obs:bucket:GetBucketLocation",
                "obs:object:GetObjectVersionAcl",
                "obs:bucket:GetBucketAcl",
                "obs:object:ListMultipartUploadParts",
                "obs:bucket:ListBucketVersions",
                "obs:object:GetObjectVersion",
                "obs:object:GetObjectAcl",
                "obs:bucket:GetBucketPolicy"
            ]
        }
    ]
}
```

ModelArts 自定义策略样例

针对当前帐号下创建的IAM用户，默认具备ModelArts的所有操作权限，当需要对IAM用户进行权限控制，如拒绝某用户使用某功能的权限时，可参考如下样例进行配置。

- 示例：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 FullAccess的系统策略，但不希望用户拥有ModelArts FullAccess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 FullAccess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"
              ]
        }
    ]
}
```

- 示例：拒绝用户使用开发环境功能

  此用户的策略配置示例如下所示：

```
{
   "Version": "1.1",
   "Statement": [

      {
          "Effect": "Deny",
          "Action": [
              "modelarts:notebook:list",
              "modelarts:notebook:create" ,
              "modelarts:notebook:get" ,
              "modelarts:notebook:update" ,
              "modelarts:notebook:delete" ,
              "modelarts:notebook:action" ,
              "modelarts:notebook:access"
          ]
      }
   ]
}
```

7 审计日志

7.1 支持云审计的关键操作

公有云平台提供了云审计服务。通过云审计服务，您可以记录与ModelArts相关的操作事件，便于日后的查询、审计和回溯。

前提条件

已开通云审计服务。

数据管理支持审计的关键操作列表

表 7-1 数据管理支持审计的关键操作列表

操作名称	资源类型	事件名称
创建数据集	dataset	createDataset
删除数据集	dataset	deleteDataset
更新数据集	dataset	updateDataset
发布数据集版本	dataset	publishDatasetVersion
删除数据集版本	dataset	deleteDatasetVersion
同步数据源	dataset	syncDataSource
导出数据集	dataset	exportDataFromDataset
创建自动标注任务	dataset	createAutoLabelingTask
创建自动分组任务	dataset	createAutoGroupingTask
创建自动部署任务	dataset	createAutoDeployTask
导入样本到数据集	dataset	importSamplesToDataset
创建数据集标签	dataset	createLabel

操作名称	资源类型	事件名称
更新数据集标签	dataset	updateLabel
删除数据集标签	dataset	deleteLabel
删除数据集标签和对应的样本	dataset	deleteLabelWithSamples
添加样本	dataset	uploadSamples
删除样本	dataset	deleteSamples
停止自动标注任务	dataset	stopTask
创建团队标注任务	dataset	createWorkforceTask
删除团队标注任务	dataset	deleteWorkforceTask
启动团队标注验收的任务	dataset	startWorkforceSamplingTask
通过/驳回/取消验收任务	dataset	updateWorkforceSamplingTask
提交验收任务的样本评审意见	dataset	acceptSamples
给样本添加标签	dataset	updateSamples
发送邮件给团队标注任务的成员	dataset	sendEmails
接口人启动团队标注任务	dataset	startWorkforceTask
更新团队标注任务	dataset	updateWorkforceTask
给团队标注样本添加标签	dataset	updateWorkforceTaskSamples
团队标注审核	dataset	reviewSamples
创建标注成员	workforce	createWorker
更新标注成员	workforce	updateWorker
删除标注成员	workforce	deleteWorker
批量删除标注成员	workforce	batchDeleteWorker
创建标注团队	workforce	createWorkforce
更新标注团队	workforce	updateWorkforce
删除标注团队	workforce	deleteWorkforce
自动创建IAM委托	IAM	createAgency
标注成员登录labelConsole标注平台	labelConsoleWorker	workerLoginLabelConsole

操作名称	资源类型	事件名称
标注成员登出labelConsole标注平台	labelConsoleWorker	workerLogOutLabelConsole
标注成员修改labelConsole平台密码	labelConsoleWorker	workerChangePassword
标注成员忘记labelConsole平台密码	labelConsoleWorker	workerForgetPassword
标注成员通过url重置labelConsole标注密码	labelConsoleWorker	workerResetPassword

开发环境支持审计的关键操作列表

表 7-2 开发环境支持审计的关键操作列表

操作名称	资源类型	事件名称
创建Notebook	Notebook	createNotebook
删除Notebook	Notebook	deleteNotebook
打开Notebook	Notebook	openNotebook
启动Notebook	Notebook	startNotebook
停止Notebook	Notebook	stopNotebook
更新Notebook	Notebook	updateNotebook
删除NotebookApp	NotebookApp	deleteNotebookApp
切换CodeLab规格	NotebookApp	updateNotebookApp

训练作业支持审计的关键操作列表

表 7-3 训练作业支持审计的关键操作列表

操作名称	资源类型	事件名称
创建训练作业	ModelArtsTrainJob	createModelArtsTrainJob
创建训练作业版本	ModelArtsTrainJob	createModelArtsTrainVersion
停止训练作业	ModelArtsTrainJob	stopModelArtsTrainVersion
更新训练作业描述	ModelArtsTrainJob	updateModelArtsTrainDesc

操作名称	资源类型	事件名称
删除训练作业版本	ModelArtsTrainJob	deleteModelArtsTrainVersion
删除训练作业	ModelArtsTrainJob	deleteModelArtsTrainJob
创建训练作业参数	ModelArtsTrainConfig	createModelArtsTrainConfig
更新训练作业参数	ModelArtsTrainConfig	updateModelArtsTrainConfig
删除训练作业参数	ModelArtsTrainConfig	deleteModelArtsTrainConfig
创建可视化作业	ModelArtsTensorboardJob	createModelArtsTensorboardJob
删除可视化作业	ModelArtsTensorboardJob	deleteModelArtsTensorboardJob
更新可视化作业描述	ModelArtsTensorboardJob	updateModelArtsTensorboardDesc
停止可视化作业	ModelArtsTensorboardJob	stopModelArtsTensorboardJob
重启可视化作业	ModelArtsTensorboardJob	restartModelArtsgTensorboardJob

AI 应用管理支持审计的关键操作列表

表 7-4 AI 应用管理支持审计的关键操作列表

操作名称	资源类型	事件名称
创建AI应用	model	addModel
更新AI应用	model	updateModel
删除AI应用	model	deleteModel
添加转换任务	convert	addConvert
更新转换任务	convert	updateConvert
删除转换任务	convert	deleteConvert

服务管理支持审计的关键操作列表

表 7-5 服务管理支持审计的关键操作列表

操作名称	资源类型	事件名称
部署服务	service	addService
删除服务	service	deleteService
更新服务	service	updateService
启停服务	service	startOrStopService
启停边缘服务节点	service	startOrStopNodesService
添加用户访问秘钥	service	addAkSk
删除用户访问秘钥	service	deleteAkSk
创建专属资源池	cluster	createCluster
删除专属资源池	cluster	deleteCluster
添加专属资源池节点	cluster	addClusterNode
删除专属资源池节点	cluster	deleteClusterNode
获取专属资源池创建结果	cluster	createClusterResult

AI Gallery 支持审计的关键操作列表

表 7-6 AI Gallery 支持审计的关键操作列表

操作名称	资源类型	事件名称
发布资产	ModelArts_Market	create_content
修改资产信息	ModelArts_Market	modify_content
发布资产新版本	ModelArts_Market	add_version
订阅资产	ModelArts_Market	subscription_content
收藏资产	ModelArts_Market	star_content
取消收藏资产	ModelArts_Market	cancel_star_content
点赞资产	ModelArts_Market	like_content
取消点赞资产	ModelArts_Market	cancel_like_content
发布实践	ModelArts_Market	publish_activity
报名实践	ModelArts_Market	regist_activity
修改个人资料	ModelArts_Market	update_user

7.2 查看审计日志

在您开启了云审计服务后，系统会记录ModelArts的相关操作，且控制台保存最近7天的操作记录。本节介绍如何在云审计服务管理控制台查看最近7天的操作记录。

操作步骤

1. 登录云审计服务管理控制台。
2. 在管理控制台左上角单击 图标，选择区域。
3. 在左侧导航栏中，单击“事件列表”，进入“事件列表”页面。
4. 事件列表支持通过筛选来查询对应的操作事件。当前事件列表支持四个维度的组合查询，详细信息如下：
   - 事件来源、资源类型和筛选类型。

     在下拉框中选择查询条件。

     其中筛选类型选择事件名称时，还需选择某个具体的事件名称。

     选择资源ID时，还需输入某个具体的资源ID。

     选择资源名称时，还需选择或手动输入某个具体的资源名称。
   - 操作用户：在下拉框中选择某一具体的操作用户，此操作用户指用户级别，而非租户级别。
   - 事件级别：可选项为“所有事件级别”、“normal”、“warning”、“incident”，只可选择其中一项。
   - 时间范围：可选择查询最近七天内任意时间段的操作事件。
5. 在需要查看的事件左侧，单击 展开该事件的详细信息。
6. 单击需要查看的事件“操作”列的“查看事件”，可以在弹窗中查看该操作事件结构的详细信息。

   更多关于云审计服务事件结构的信息，请参见《云审计服务用户指南》。

8 建议反馈

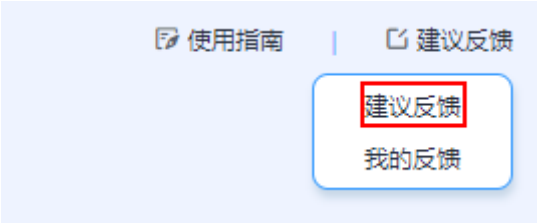
ModelArts建议反馈功能旨在收集、跟踪、解决用户问题。当前ModelArts提供一种实时的截图发帖反馈方式，帮助用户高效、快捷地进行建议反馈。ModelArts建议反馈功能仅在华为云中国站中文上使用。

在 Chrome 安装插件

ModelArts建议反馈功能需要插件支持。当前建议反馈插件仅支持Chrome浏览器，建议您在Chrome浏览器环境下完成插件的安装。

1. 登录ModelArts控制台，进入任意业务页面，单击页面右上角的“建议反馈>建议反馈”。

   图 8-1 建议反馈

2. 当您未安装ModelArts建议反馈插件时，弹窗存在提醒，如图8-2所示。单击图8-2中的链接可跳转至资料页面，资料页面提供插件下载地址。单击下载地址，即可完成ModelArts建议反馈插件的下载。

   图 8-2 建议反馈提醒

3. 解压ModelArts建议反馈插件压缩文件modelarts-screenshot-feedback.zip至任意路径，得到modelarts-screenshot-feedback文件夹。
4. 打开Chrome浏览器，单击Chrome浏览器的设置-扩展程序选项，进入扩展程序页面。
5. 在扩展页面中，打开右上角的开发者模式开关，如图8-3所示。

图 8-3 开发者模式开关

6. 拖动已解压的ModelArts建议反馈插件文件夹modelarts-screenshot-feedback到Chrome浏览器扩展程序页面，即可完成插件的安装。

截图发帖反馈

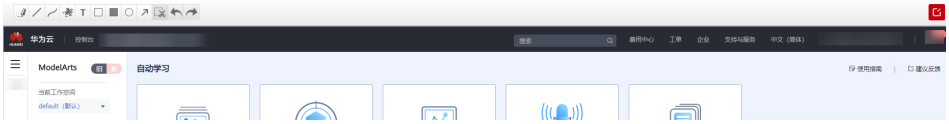
ModelArts建议反馈功能提供实时截图能力并支持截图编辑操作。

1. 登录ModelArts平台，进入任意业务页面，单击右上角“建议反馈>建议反馈”。
2. 在弹窗页面选择“是”进入页面截图编辑状态。您也可以选择“否”直接进行建议发帖。

图 8-4 截图反馈确认

3. 进入当前页面截图可编辑状态。用户可以通过左侧的图形编辑功能进行截图的编辑。

图 8-5 建议反馈可编辑状态页面



插件功能的图标从左到右分别为：

- 高亮
- 直线/曲线
- 喷枪

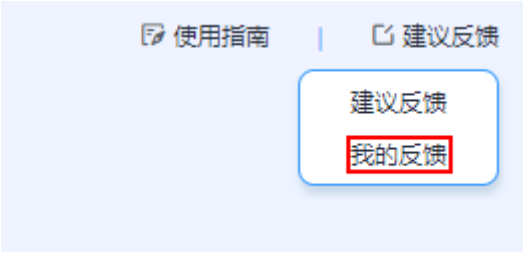
- 文字
- 矩形框/矩形方块/圆框
- 箭头
- 裁剪
- 撤销
- 恢复
- 发帖

4. 完成截图编辑后，用户可以通过单击右侧发帖反馈键跳转至图 发帖界面进行帖子的编辑和发布。

进入我的反馈

1. 登录ModelArts控制台，进入任意业务页面，单击页面右上角“建议反馈>我的反馈”。

图 8-6 我的反馈



2. 页面跳转至“个人中心>我的论坛”。您可以在此页面跟踪您的建议反馈进展。您可以根据实际情况查看您的“最新发布”和“最新回复”。

图 8-7 我的反馈详细页

A 修订记录

发布日期	修改说明
2023-08-30	下线开发环境（旧版）资料。
2023-03-01	下线免费体验ModelArts章节，内容迁移至《准备工作》的“开通ModelArts资源 > **免费体验**”专题。 下线审计日志章节，内容迁移至《资源管理》的**审计日志**专题。 下线权限管理章节，内容迁移至《最佳实践》的**权限管理**专题。 下线AI工程师如何使用ModelArts章节。
2022-02-30	下线AI应用管理、部署服务、模型包规范、模型版本、自定义脚本代码示例、监控章节。内容迁移至《**推理部署**》。
2022-05-19	推理服务支持无损滚动升级。
2022-01-26	下线部署服务章节的“难例筛选”。
2022-01-04	下线数据管理章节中的“一键模型”上线功能。
2021-12-20	旧版训练中下线预置算法。 下线模型评估章节，转为邀测特性。
2021-12-15	体验优化：“模型管理”更名为“AI应用”。
2021-11-25	资源池中优化部分描述内容。
2021-10-21	新增11.7-旧版训练与新版训练差异
2021-9-27	更新开发环境New中的VSCode插件下载地址。 新增开发环境New中的变更Notebook实例运行规格。
2021-09-18	新增管理数据（New），数据集与数据标注任务解耦。 新增配置SSH工具远程连接开发环境Notebook。 新增配置VSCode插件远程连接开发环境Notebook 新增开发环境New中的动态挂载OBS并行文件系统。
2021-08-30	开发环境（New）中优化部分描述内容。

发布日期	修改说明
2021-07-17	新增发布开发环境（New）。
2021-06-26	新增发布训练管理（New）。
2021-04-23	新增发布CodeLab。
2021-03-26	• 在线服务新增个性化配置。 • “图像分割”类型数据集支持从OBS导入数据
2021-01-14	• 智能标注功能优化。 • AI市场更名。 • 模型包规范新增机器学习推理代码样例和配置文件说明。 • 自定义脚本代码示例优化Pytorch代码示例。
2020-12-10	• 团队标注功能优化。 • 创建数据集时，随软件界面更新。 创建数据集（旧版） • 创建或者扩容推理专属资源池支持指定可用区。 创建专属资源池
2020-11-10	• 创建Notebook时，优化工作环境的展示。 8.2.1-创建Notebook实例 • 增加模型评估诊断后的优化建议。 • 新增数据处理功能。
2020-10-15	• 新增建议反馈功能。 建议反馈 • 在线服务新增AK/SK认证介绍。
2020-09-21	• 在线服务新增支APP认证方式的功能。
2020-09-11	• 新增支持模型评估诊断的功能。 • 查看训练作业时，支持查看评估结果。 评估结果 添加评估结果
2020-08-06	• Notebook功能描述优化。重新优化大纲结构，同时增加JupyterLab常用操作指导。 8.3.1-Jupyter Notebook简介 8.4.1-JupyterLab简介及常用操作 • “表格”类型数据集，支持DWS数据源输入。 表格 • 部署在线服务API增加集成API的指导。 • 旧版“AI市场”及“数据管理”下线。

发布日期	修改说明
2020-07-06	数据集管理新增支持视频数据。 自动化搜索作业优化增强，同时增加1个示例。 开发环境增加支持GitHub代码库功能。 ● 8.2.2-创建带有Git存储库的Notebook实例 ● 8.4.6-使用Git插件’ 开发环境、模型导入支持PyTorch 1.4.0引擎。 ● 8.2.1-创建Notebook实例 增加TensorFlow 2.X的自定义脚本代码示例。 模型转换任务创建页面优化。
2020-06-08	数据集管理新增支持表格型数据，支持导入和发布csv。 数据特征分析，增加特征分析任务启动和管理功能。 开发环境增强，引入支持JupyterLab、引擎新增TF2.1，对ModelArts多种基础能力的集成以及若干体验改进。 ● 8.2.1-创建Notebook实例 ● 8.3.5-使用ModelArts示例 ● 8.3.7-与OBS同步文件 ● 8.4.1-JupyterLab简介及常用操作 ● 8.3.4.1-上传大文件至Notebook中 ● 8.3.4.2-下载Notebook内的大文件到本地 自动化搜索作业优化：新增fix_norm超参搜索、adv_aug数据增强、betanas等多种NAS方法，达到ImageNet上mobile setting下最佳精度，支持最低5行代码即可享用多元搜索能力。 预置框架支持TenorFlow 2.1。创建Notebook实例或使用常用框架创建训练作业时，可选择TenorFlow 2.1。 ● **使用常用框架训练模型** ● 8.1-Notebook简介 模型管理新增支持python3.7的runtime环境。
2020-04-07	● 支持Ascend训练，AI市场提供了可用于Ascend训练的云端算法。 ● 部署上线，开放支持Ascend推理能力。
2020-03-25	● 提供限时免费的资源规格，可在自动学习、训练作业以及Notebook中使用，增加相应说明。 – **免费体验自动学习** – **免费体验AI全流程开发** – 免费体验Notebook ● 在AI市场，ModelArts官方提供了云端算法，无需编码，快速启动模型训练。

发布日期	修改说明
2020-03-20	● 优化训练作业相关说明，按不同算法来源进行描述。 – 使用已有算法训练模型 – 使用常用框架训练模型 – 使用自定义镜像训练模型 ● 文本三元组、文本分类、命名实体类型的数据集支持团队标注，刷新数据标注功能说明。 ● ModelArts UI风格升级，刷新本文档所有相关截图。
2020-02-21	● 新增两个的转换模板。
2020-01-07	● 旧版数据管理功能隐藏。
2019-12-03	● 对模型导入功能，按场景拆分为4个子章节。 ● 针对支持团队标注的数据集，支持管理团队标注任务。 ● 增加数据集类型“文本三元组” ● 数据管理的整体章节进行编辑优化。 ● 增加自动化搜索作业管理的功能。
2019-11-06	● 增加模型订阅管理功能。 ● 部署边缘服务时，在计算节点规格中，增加“自定义规格”。
2019-10-17	● 数据管理：随软件变更，优化并新增功能。所有相关章节刷新描述，并且新增如下章节。 ● 增加自定义脚本代码示例（包含常用引擎）。 ● 增加监控相关描述
2019-09-30	● 对模型模板的描述进行丰富和优化。 ● 优化模型包规范的描述，并丰富了模型包示例。 ● 新增提供模型转换功能。
2019-08-08	● 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。
2019-07-18	修改Notebook中选择AI引擎的方式。修改如下章节： 8.1-Notebook简介 8.2.1-创建Notebook实例
2019-06-30	新增Notebook上传大文件的功能。 8.3.4.1-上传大文件至Notebook中
2019-06-20	新增关于模型模板的输入输出模式说明。
2019-06-03	导入模板时，新增从模板中导入的功能。 开发环境中，增加“Multi-Engine(Recommend)”引擎。 ● 修改：8.1-Notebook简介、8.2.1-创建Notebook实例

发布日期	修改说明
2019-05-31	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。 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。 本次刷新，修改了所有大纲，并优化了每个章节的描述语言。
2019-04-16	修改 ● 准备工作章节内容。 ● 开发环境章节内容。
2019-04-12	新增 ● 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。 修改 ● 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	新增 ● 数据标注-声音分类章节内容。
2019-04-01	修改 ● 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	新增 ● 新增数据标注-Beta章节内容。
2019-03-18	修改 ● 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	新增 ● 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	新增 ● 新增了模型包规范。 ● 新增Notebook中8.3.7-与OBS同步文件介绍、8.3.6-使用ModelArts SDK等内容。
2018-12-21	修改 ● 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。
2018-12-03	修改 ● 调整第三章节结构、修改部分内容。
2018-11-15	修改 ● 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。