ModelArts
3.3.5
数据准备与分析用户指南

文档版本 03
发布日期 2022-03-24

华为技术有限公司

商标声明


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。

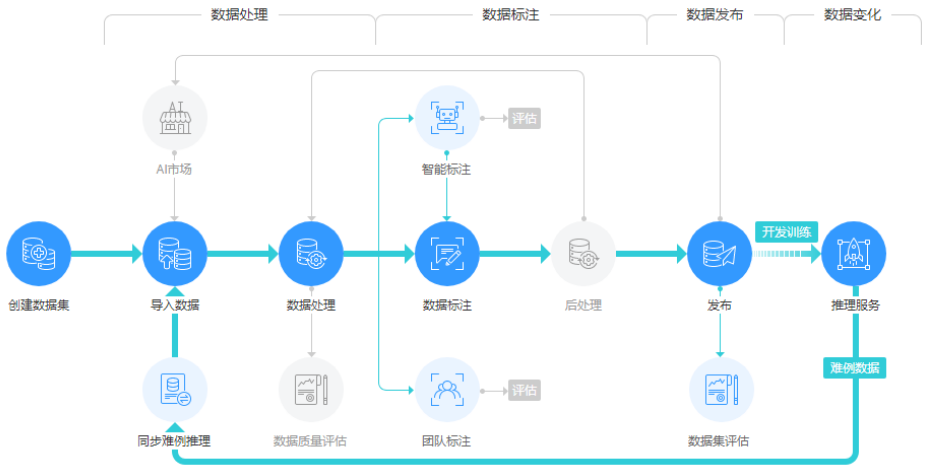
目 录

1 数据准备简介

通常来讲，AI人工智能三要素包括数据、算法和算力。数据的质量会影响模型的精度，一般来说，大量高质量的数据更有可能训练出高精度AI模型。现在很多算法使用常规数据能将准确率做到85%或者90%，而商业化应用往往要求更高，如果将要模型精度提升至96%甚至99%，则需要大量高质量的数据，这个时候也会要求数据更加精细化、场景化、专业化，这往往也成为了AI模型突破瓶颈的关键性条件。如何快速准备大量高质量的数据已经成为AI开发过程中一个极具挑战性的问题。

ModelArts是面向AI开发者的一站式开发平台，能够支撑开发者从数据到AI应用的全流程开发过程，包含数据处理、算法开发、模型训练、模型部署等操作。并且提供AI Gallery功能，能够在市场内与其他开发者分享数据、算法、模型等。为了能帮用户快速准备大量高质量的数据，ModelArts数据管理提供了全流程的数据准备、数据处理和数据标注能力。

图 1-1 ModelArts 数据准备全流程



ModelArts数据管理为用户准备高质量的AI数据提供了以下主要能力：

- 解决用户获取数据的问题。
  - 用户可在AI Gallery上一键下载需要的数据资源到ModelArts数据管理。
  - 提供多种数据接入方式，支持用户从OBS，MRS，DLI以及DWS等服务导入用户的数据。

  - 提供18+数据增强算子，帮助用户扩增数据，增加训练用的数据量。
- 帮助用户提高数据的质量。
  - 提供图像、文本、音频、视频等多种格式数据的预览，帮助用户识别数据质量。
  - 提供对数据进行多维筛选的能力，用户可以根据样本属性、标注信息等进行样本筛选。
  - 提供12+标注工具，方便用户进行精细化、场景化和专业化的数据标注。
  - 提供基于样本和标注结果进行特征分析，帮助用户整体了解数据的质量。
- 提升用户数据准备的效率。
  - 提供数据版本管理能力，帮助用户提升数据管理的效率。
  - 提供数据校验、数据选择、数据清洗等多种数据处理算子，帮助用户快速处理数据。
  - 提供交互式标注、智能标注等能力，提升用户数据标注的效率。
  - 提供团队标注以及团队标注流程管理能力，帮助用户提升大批量数据标注的能力。

2 入门教程

本节以准备训练物体检测模型的数据为例，介绍如何针对样例数据，进行数据分析、数据标注等操作，完成数据准备工作。在实际业务开发过程中，可以根据业务需求选择数据管理的一种或多种功能完成数据准备。此次操作分为以下流程：

- **准备工作**
- **创建数据集**
- **数据分析**
- **数据标注**
- **数据发布**
- **数据导出**

准备工作

在使用ModelArts数据管理的功能前，需要先完成以下准备工作。

用户在使用数据管理的过程中，ModelArts需要访问用户的OBS等依赖服务，需要用户进行在“全局配置”页面中进行委托授权。具体操作参考**使用委托授权（推荐）**。

创建数据集

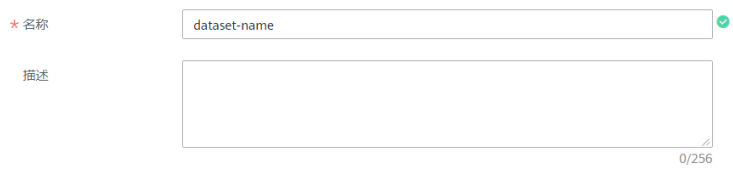
本示例使用OBS中的数据作为数据集的输入目录创建数据集。参考如下操作创建一个物体检测类型的数据集，并将数据导入到数据集中。

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。

步骤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。

1. 填写数据集基本信息，数据集的“名称”和“描述”。

图 2-1 数据集基本信息

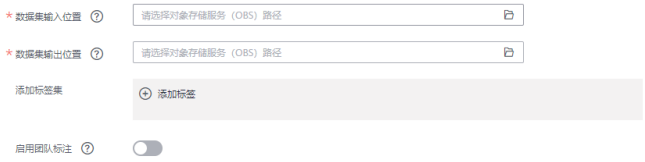


2. 选择“标注场景”和“标注类型”，本案例中分别选择“图片”和“物体检测”。

图 2-2 数据集标注场景和标注类型

3. 选择OBS中的数据目录作为“数据集输入位置”，选择不同的OBS目录作为“数据集输出位置”。

图 2-3 数据集的输入位置和输出位置



4. 参数填写无误后，单击页面右下角“创建”，即可完成数据集的创建。

----结束

数据分析

数据集创建完成后，可以基于图片各项特征，如模糊度、亮度等进行分析，帮助用户更好的分析数据集的数据质量，判断数据集是否满足自己的算法和模型要求。

1. 创建特征分析任务

   a. 在执行特征分析前，需先发布一个数据集版本。在“数据集概览”页单击右上角的“发布”，为数据集发布一个新版本。

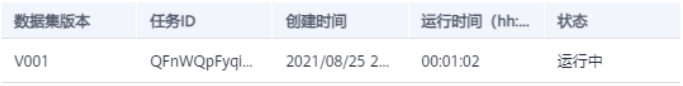
   b. 版本发布完成后，进入数据集概览页。选择“数据特征”页签，单击“特征分析”，在弹窗中选择刚才发布的数据集版本，并单击“确定”，启动特征分析任务。

图 2-4 启动特征分析

c. 查看任务进度

任务执行过程中，可以单击“任务历史”，查看任务进度。当任务状态变为“成功”时，表示任务执行完成。

图 2-5 特征分析任务进度

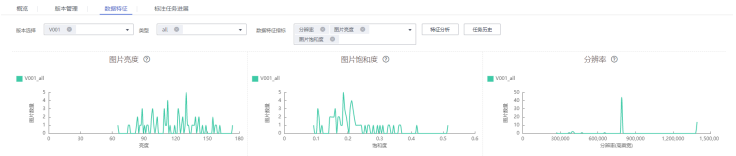


2. 查看特征分析结果

特征分析任务执行完成后，可以在“数据特征”页签下，选择“数据集版本”、“类型”和“数据特征指标”，页面将自动呈现您选择对应版本及其指标数据，您可以根据呈现的图表了解数据分布情况，帮助您更好的理解您的数据。

- “版本选择”：根据实际情况选择已执行过特征任务的版本，可以选多个进行对比，也可以只选择一个。
- “类型”：根据需要分析的类型选择。支持“all”、“train”、“eval”和“inference”。分别表示所有、训练、评估和推理类型。
- “数据特征指标”：选择您需要展示的指标。详细指标解释，可参见【**特征分析指标列表**】。

图 2-6 查看特征分析结果



在特征分析结果中，例如图片亮度指标，数据分布中，分布不均匀，缺少某一种亮度的图片，而此指标对模型训练非常关键。此时可选择增加对应亮度的图片，让数据更均衡，为后续模型构建做准备。

数据标注

- **人工标注**

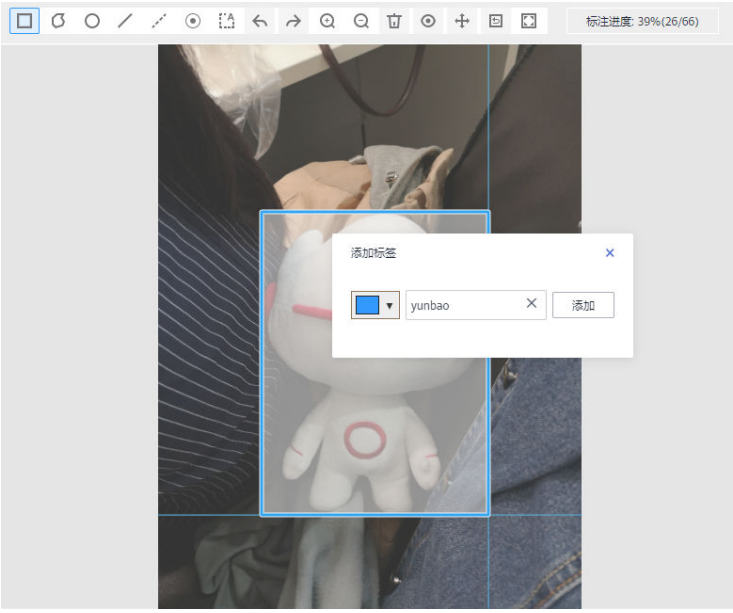
a. 在“未标注”页签图片列表中，单击图片，自动跳转到标注页面。

b. 在标注页面的工具栏中选择合适的标注工具，本示例使用矩形框进行标注。

图 2-7 标注工具

c. 使用标注工具选中目标区域，在弹出的标签文本框中，直接输入新的标签名。如果已存在标签，从下拉列表中选择已有的标签。单击“添加”完成标注。

图 2-8 添加物体检测标签



d. 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。

- **智能标注**

  通过人工标注完成少量数据标注后，可以通过智能标注对剩下的数据进行自动标注，提高标注的效率。

  a. 在数据集详情页面，单击右上角“启动智能标注”。

  b. 在“启动智能标注”窗口中，填写如下参数，然后单击“提交”。

     - **“智能标注类型“**：主动学习
     - **“算法类型“**：快速型

     其他参数采用默认值。

图 2-9 启动智能标注任务

c. 查看智能标注任务进度

智能标注任务启动后，可以在“待确认”页签下查看智能标注任务进度。当任务完成后，即可在“待确认”页签下查看自动标注好的数据。

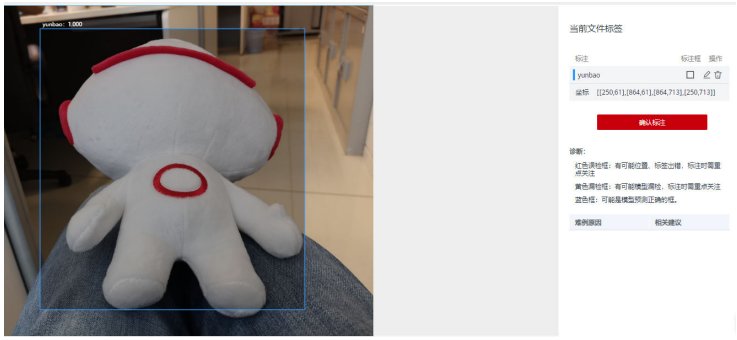
图 2-10 查看智能标注任务进度

d. 确认智能标注结果

在智能标注任务完成后，在“待确认”页签下，单击具体图片进入标注详情页面，可以查看或修改智能标注的结果。

如果智能标注的数据无误，可单击右侧的“确认标注”完成标注，如果标注信息有误，可直接删除错误标注框，然后重新标注，以纠正标注信息。针对物体检测任务，需一张一张确认。确保所有图片已完成确认，然后执行下一步操作。

图 2-11 确认智能标注结果



数据发布

ModelArts训练管理模块支持通过ModelArts数据集或者OBS目录中的文件创建训练作业。如果选择通过数据集作为训练任务的数据源，则需要指定数据集及特定的版本。因此，用户需要为准备好的数据发布一个版本，具体操作参考**发布数据版本**。

图 2-12 创建训练任务的数据来源

数据导出

ModelArts训练管理模块支持通过ModelArts数据集或者OBS目录中的文件创建训练作业。如果选择通过OBS目录的方式创建训练任务，用户需要将数据集中准备好的数据导出到OBS中。

1. 导出数据到OBS
   a. 在数据集详情页面中，选中需要导出的数据或筛选出需要导出的数据，然后单击右上角“导出”。
   b. 导出方式选择“OBS”，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。

      图 2-13 导出至 OBS

   c. 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。
2. 查看任务历史
   当您导出数据后，可以通过任务历史查看导出任务明细。
   a. 在数据集详情页面中，单击右上角“任务历史”。
   b. 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 2-14 导出任务历史

3 创建数据集

在ModelArts进行数据准备，首先需要先创建一个数据集，后续的操作如数据导入、数据分析、数据标注等，都是基于数据集来进行的。

3.1 数据集简介

数据集的类型

当前ModelArts支持如下格式的数据集，包含文件型和表格型。

- 文件型
  - 图片：对图像类数据进行处理，支持 .jpg、.png、.jpeg、.bmp四种图像格式，支持用户进行图像分类、物体检测、图像分割类型的标注。
  - 音频：对音频类数据进行处理，支持.wav格式，支持用户进行声音分类、语音内容、语音分割三种类型的标注。
  - 文本：对文本类数据进行处理，支持.txt、.csv格式，支持用户进行文本分类、命名实体、文本三元组三种类型的标注。
  - 视频：对视频类数据进行处理，支持.mp4格式，支持用户进行视频标注。
  - 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 3-1 自由格式数据集示例

- 表格型

表格：适合表格等结构化数据处理。数据格式支持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。

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，如智能标注、团队标注等。详细信息参考**表3-1**。

表 3-1 不同类型的数据集支持的功能

数据集类型	标注类型	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	自动分组	数据特征
文件型	图像分类	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
	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	-	-
	文本分类	支持	支持	-	支持	支持	支持	-	-
	命名实体	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	-	-
	视频	支持	支持	-	支持	支持	支持	-	-
	自由格式	支持	-	_	支持	支持	支持	-	-
表格型	表格	支持	支持	-	支持	支持	支持	-	-

规格限制

- 除表格类型之外的数据集（如视频、文本、音频等），单个数据集的最大样本数量限制：1000000，最大标签数量限制：10000。

- 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。

3.2 创建数据集

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

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。

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

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。当前不支持OBS并行文件系统，请选择OBS对象存储。

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”，单击“前往新版”，进入新版“数据集”管理页面。

   图 3-2 进入新版数据集

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”、“描述”、“数据格式”、“数据类型”和“数据集输出位置”。ModelArts目前支持的类型及其说明请参见**数据集的类型**。

图 3-3 数据集基本信息

b. 基本信息填写完成后，单击“下一步”，填写数据集的数据输入信息。

i. OBS导入数据

用户在OBS中有准备好的数据时，选择“OBS”，“导入路径”、“数据标注状态”、和数据“标注格式”。针对不同类型的数据集，数据输入支持的标注格式不同，ModelArts目前支持的标注格式及其说明请参见**数据接入简介**。

图 3-4 选择 OBS 中的数据格式和数据类型

ii. 从AI Gallery下载数据

当用户没有准备数据时，可以从AI Gallery上下载数据创建数据集。选择“AI Gallery”并选中列表中的一个资产，在“下载至OBS桶位置（数据集输入位置）”选择一个空目录用来存储下载的数据集。

图 3-5 AI Gallery 下载数据

c. 参数填写无误后，单击页面右下角“提交”。

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，您可以执行数据导入、发布、修改、删除、数据处理、数据标注、数据特征、版本管理和导出操作。

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**。

文件型（图片、音频、文本、视频、自由格式）

表 3-2 数据集的详细参数

参数名称	说明
数据集输出位置	选择数据集输出位置的OBS路径，此位置会存放输出的标注信息等文件。 **说明** “数据集输出位置”不能与“数据输入路径”为同一路径，且不能是“数据输入路径”的子目录。 “数据集输出位置”建议选择一个空目录。 “数据集输出位置”不支持OBS并行文件系统下的路径，请选择OBS对象桶。
导入路径	选择需要导入数据的OBS路径，此位置会作为数据集的数据存储路径。 **说明** “导入路径”不支持OBS并行文件系统下的路径，请选择OBS对象桶。 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。如果需要在数据集中修改数据，建议使用**同步数据源**或4章节**OBS目录导入操作**功能。 超出数据集的样本和标签配额，会导致数据无法正常导入。

参数名称	说明
数据标注状态	选择数据的标注状态，分为“未标注”和“已标注”。 选择“已标注”时，需指定标注格式，并保证数据文件满足相应的格式规范，否则可能存在导入失败的情况。 仅图片（物体检测、图像分类、图像分割）、音频（声音分类）、文本（文本分类）类型的标注任务支持导入已标注数据。

表格型（表格）

图 3-6 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 3-3 数据集的详细参数

参数名称	说明
数据集输出位置	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。
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**》。

参数名称	说明
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3-4。 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 3-4 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

3.3 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“修改”。
3. 参考表3-5修改数据集基本信息，然后单击“确定”完成修改。

图 3-7 修改数据集

表 3-5 参数说明

参数	说明
名称	数据集的名称，支持1~64位可见字符。名称只能是字母、中文、数字、下划线或者中划线组成的合法字符串。且只能以字母或者中文字符开头。
描述	数据集的简要描述。

4 数据接入

4.1 数据接入简介

数据集创建完成后，您还可以通过导入数据的操作，接入更多数据。ModelArts支持从不同数据源导入数据。

- **从AI Gallery下载数据集**
- **从OBS导入数据**
- **从DLI导入数据**
- **从MRS导入数据**
- **从DWS导入数据**
- **从本地上传数据**

ModelArts的AIGallery中预置了大量的数据集，文件型和表格型都有，用户可以从AIGallery下载使用预置的数据集。您也可以将您自己的数据导入到ModelArts中。

文件型数据来源

除了从AIGallery下载预置数据集外，文件型数据集还支持从两种数据源导入数据：“OBS”和“本地上传”。导入后，导入目录下的数据会拷贝至数据集的数据源路径下。

- OBS：又分为从OBS目录或从Manifest文件两种导入方式，需要将导入的数据或Manifest文件提前存储至OBS目录中。
- 本地上传：将本地数据直接通过Internet上传至OBS指定目录后，再导入数据集。

表格型数据来源

除了从AIGallery下载预置数据集外，表格数据集还支持从5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）和本地上传。

数据集中的数据导入入口

数据集中的数据导入有5个入口。

- 创建数据集时直接从设置的数据导入路径中自动同步数据。

图 4-1 创建数据集时导入数据

- 创建完数据集后，在数据集列表页面的操作栏单击“导入”，导入数据。

图 4-2 在数据集列表页导入数据

- 在数据集列表页面，单击某个数据集的名称，进入数据集详情页中，单击“导入”，导入数据。

图 4-3 在数据集详情页中导入数据

- 在数据集列表页面，单击某个数据集的名称，进入数据集详情页中，单击“同步数据源”，同步OBS中的数据。

图 4-4 在数据集详情页中同步数据源

- 在数据标注的标注作业详情中添加数据。

图 4-5 标注作业详情中添加数据

4.2 从 AI Gallery 下载数据集

ModelArts 的AI Gallery提供了丰富的数据资源，用户可以查找并下载满足业务需要的数据集，直接用于创建训练作业。

1. 登录ModelArts管理控制台，在左侧菜单栏中选择“AI Gallery”，进入AI Gallery页面。
2. 选择“资产集市 > 数据”页签，默认进入公共页面，该页面展示了所有共享的数据集。

图 4-6 AI Gallery 数据页面

3. 搜索业务需要的数据集，单击目标数据集进入详情页面。
4. 在数据集详情页面单击“下载”。选择数据集下载方式，下载至OBS或者ModelArts数据集。
   a. 将数据集下载至OBS
      - “下载方式”选择“对象存储服务（OBS）”。
      - “目标区域”选择您需要将该数据集下载到的区域位置，如“华北-北京四”。
      - “目标位置”选择OBS桶路径，桶内如有同名的文件或文件夹，将被新下载的文件或文件夹覆盖。

图 4-7 下载数据集到 OBS

   b. 将数据集下载至ModelArts
      - “下载方式”选择“ModelArts数据集”。
      - “目标区域”选择您需要将该数据集下载到的区域位置，如“华北-北京四”。
      - “目标位置”选择OBS桶路径，桶内如有同名的文件或文件夹，将被新下载的文件或文件夹覆盖。

- “名称”默认生成“data-xxxx”形式的数据集名称，该数据集会同步在ModelArts数据集列表中。
- “描述”可以添加对于该数据集的相关描述。

图 4-8 下载数据集到 ModelArts

5. 单击“确定”，自动跳转至“个人中心 > 我的数据 > 我的下载”页面，查看下载详情。下载成功后，数据集详情列表会显示“文件大小”。

4.3 从 OBS 导入数据

4.3.1 OBS 导入数据简介

导入方式

OBS导入数据方式分为“OBS目录”和“Manifest文件”两种。

- OBS目录：指需要导入的数据集已提前存储至OBS目录中。此时需选择用户具备权限的OBS路径，且OBS路径内的目录结构需满足规范，详细规范请参见**OBS目录导入数据规范说明**。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**Manifest文件导入规范说明**。

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 4-1 不同类型数据集支持的导入方式

分类	数据集类型	标注类型	OBS目录导入	Manifest文件导入
文件型	图片	图像分类	支持 可以导入未标注或已标注数据 已标注数据格式规范：**图像分类**	支持 可以导入未标注或已标注数据 已标注数据格式规范：**图像分类**
		物体检测	支持 可以导入未标注或已标注数据 已标注数据格式规范：**物体检测**	支持 可以导入未标注或已标注数据 已标注数据格式规范：**物体检测**
		图像分割	支持 可以导入未标注或已标注数据 已标注数据格式规范：**图像分割**	支持 可以导入未标注或已标注数据 已标注数据格式规范：**图像分割**
	音频	声音分类	支持 导入的是未标注或已标注数据 格式规范：**声音分类**	支持 可以导入未标注或已标注数据 已标注数据格式规范：**声音分类**
		语音内容	支持 导入的是未标注数据	支持 可以导入未标注或已标注数据 已标注数据格式规范：**语音内容**
		语音分割	支持 导入的是未标注数据	支持 可以导入未标注或已标注数据 已标注数据格式规范：**语音分割**
	文本	文本分类	支持 导入的是未标注或已标注数据 已标注数据格式规范：**文本分类**	支持 可以导入未标注或已标注数据 已标注数据格式规范：**文本分类**

分类	数据集类型	标注类型	OBS目录导入	Manifest文件导入
		命名实体	支持 导入的是未标注数据	支持 可以导入未标注或已标注数据 已标注数据格式规范：**文本命名实体**
		文本三元组	支持 导入的是未标注数据	支持 可以导入未标注或已标注数据 已标注数据格式规范：**文本三元组**
	视频	视频	支持 导入的是未标注数据	支持 可以导入未标注或已标注数据 已标注数据格式规范：**视频标注**
	其他	自由格式	支持 导入的是未标注数据	-
表格型	表格	表格	支持 还支持从DWS、DLI、MRS导入数据。 格式规范：**表格**	-

4.3.2 OBS 目录导入操作

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。详细指导请参见**如何上传数据至OBS**。
- 确保数据存储的OBS桶与ModelArts在同一区域，并确保用户具有OBS桶的操作权限。

文件型数据从 OBS 目录导入操作

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“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”。
   - “数据来源”：“OBS”
   - “导入方式”：“目录”。
   - “导入路径”：数据存储的OBS路径。
   - “数据标注状态”：已标注。
   - “高级特征选项 ”：默认关闭，可通过勾选高级选项提供增强功能。

     如“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”添加相应的标签。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。

图 4-9 导入数据集-OBS

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，并可以通过创建标注任务进行数据标注。

文件型数据标注状态

数据标注状态分为“未标注”和“已标注”。

- 未标注：仅导入标注对象（指待标注的图片，文本等），不导入标注内容（指标注结果信息）。
- 已标注：同时导入标注对象和标注内容，当前“自由格式”的数据集不支持导入标注内容。

  为了确保能够正确读取标注内容，要求用户严格按照规范存放数据：

  导入方式选择目录时，需要用户选择“标注格式”，并按照标注格式的要求存放数据，详细规范请参见**标注格式**章节。

  导入方式选择manifest时，需要满足manifest文件的规范。

  说明

  数据标注状态选择“已标注”，您需要保证目录或manifest文件满足相应的格式规范，否则可能存在导入失败的情况。

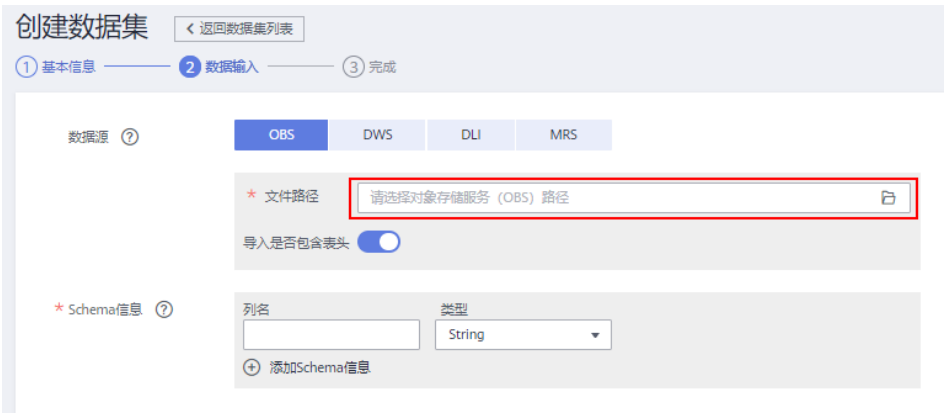
表格数据集从 OBS 导入操作

Modelarts支持从OBS导入表格数据，即csv文件。

表格数据集导入说明：

- 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。
- 从OBS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。如果数据格式不合法，会将数据置为null，详见**表3-4**。
- 导入的csv文件要求如下：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
```



4.3.3 OBS 目录导入数据规范说明

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”、“图像分割”、“文本分类”和“声音分类”标注类型支持按标注格式导入。

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：

1）ModelArts imageNet 1.0：目录方式，只支持单标签

- 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
```

2）ModelArts image classification 1.0：txt标签文件，支持多标签

- 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。

  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

支持两种格式：

1)ModelArts PASCAL VOC 1.0

- 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4-9**。

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。

2）YOLO：

- YOLO数据集目录应该遵循以下结构，格式不规范将导致导入任务失败。

```
└── yolo_dataset/
   |
   ├──── obj.names  # 标签集文件
   ├──── obj.data   # 记录数据集信息的文件及路径信息(相对路径)
   ├──── train.txt  # 训练集中各图片路径信息(相对路径)
   ├──── valid.txt  # 验证集中各图片路径信息(相对路径)
   |
   ├──── obj_train_data/  # 训练集的图片与对应的标注文件所在目录
   |     ├──── image1.txt  # image1对应的带标签bbox列表
   |     ├──── image1.jpg
   |     ├──── image2.txt
   |     ├──── image2.jpg
```



```
|    ├──── ...
|
├──── obj_valid_data/  # 验证集的图片与对应的标注文件所在目录
|    ├──── image101.txt
|    ├──── image101.jpg
|    ├──── image102.txt
|    ├──── image102.jpg
|    ├──── ...
```

YOLO数据集只支持train和valid子集。如果导入的数据集包括除了上述之外的子集，这些其他子集将被忽略。

- obj.data应包含以下内容，train和valid子集必许至少有一个，其中文件路径均为相对路径。

```
classes = 5 # 可选
names = <path/to/obj.names>#例如 obj.names
train = <path/to/train.txt>#例如 train.txt
valid = <path/to/valid.txt>#可选， 例如valid.txt
backup = backup/ # 可选
```

- obj.names文件记录标签列表。每一行的行标作为该标签的index。

```
label1  # label1的index：  0
label2  # label2的index：  1
label3
...
```

- train.txt和valid.txt文件结构如下，其中文件路径均为相对路径，且文件列表与目录下的文件需一一对应：

```
<path/to/image1.jpg>#例如 obj_train_data/image.jpg
<path/to/image2.jpg>#例如 obj_train_data/image.jpg
...
```

- obj_train_data/ 和 obj_valid_data/目录下的.txt文件应该包含对应图像的bbox标签信息，每行代表一个bbox标注。

```
# image1.txt:
# <label_index> <x_center> <y_center> <width> <height>
0 0.250000 0.400000 0.300000 0.400000
3 0.600000 0.400000 0.400000 0.266667
```

其中x_center, y_center, width, and height分别表示归一化后的目标框中心点x坐标、归一化后的目标框中心点y坐标、归一化后的目标框宽度、归一化后的目标框高度。

- 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

ModelArts image segmentation 1.0：

- 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4-5**。

  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。

- ModelArts text classfication combine 1.0：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
  ```

- ModelArts text classfication 1.0：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

ModelArts audio classfication dir 1.0：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

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
```

4.3.4 Manifest 文件导入操作

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，并确保用户具有OBS桶的操作权限。

文件型数据从 Manifest 导入操作

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。当前操作指导以图片数据集为例。

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“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”。
   - “数据来源”：“OBS”
   - “导入方式”：“manifest”。
   - “Manifest文件”：存储Manifest文件的OBS路径。
   - “数据标注状态”：已标注。
   - “高级特征选项 ”：默认关闭，可通过勾选高级选项提供增强功能。

     “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。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。

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

图 4-10 导入 mainfest 文件

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，并可以通过创建标注任务进行数据标注。

文件型数据标注状态

数据标注状态分为“未标注”和“已标注”。

- 未标注：仅导入标注对象（指待标注的图片，文本等），不导入标注内容（指标注结果信息）。
- 已标注：同时导入标注对象和标注内容，当前“自由格式”的数据集不支持导入标注内容。

  为了确保能够正确读取标注内容，要求用户严格按照规范存放数据：

  导入方式选择目录时，需要用户选择“标注格式”，并按照标注格式的要求存放数据。

  导入方式选择manifest时，需要满足manifest文件的规范，详细规范请参见**标注格式**章节。

  说明

  数据标注状态选择“已标注”，您需要保证目录或manifest文件满足相应的格式规范，否则可能存在导入失败的情况。

4.3.5 Manifest 文件导入规范说明

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 4-2 字段说明

字段	是否必选	说明
source	是	被标注对象的URI。数据来源的类型及示例请参考**表4-3**。
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4-4**。
inference-loc	否	当此文件由推理服务生成时会有该字段，表示推理输出的结果文件位置。

表 4-3 数据来源类型

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

表 4-4 annotation 对象说明

字段	是否必选	说明
type	是	标签类型。取值范围为： ● 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。
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4-2。
- “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 4-5 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。
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4-6**。 • mask_color：必选字段，标签的颜色，以RGB值表示

表 4-6 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4-2**。

文本命名实体

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

“source”、“usage”、“annotation”等参数说明与**图像分类**一致，详细说明请参见**表4-2**。

其中，property的参数解释如**表4-7**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 4-7 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4-2**。

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 4-8 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4-2。
- “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 4-9 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4-10**。

表 4-10 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4-2**。

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4-2**。
- “property”中的“@modelarts:content”参数，数据类型为“String”，表示语音内容。

语音分割

```
{
  "source":"s3://path/to/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4-2**。
- “property”的参数解释如**表4-11**所示。

表 4-11 “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4-2。
- “annotation-loc”：对于物体检测、是必选字段，对于其他类型是可选字段，标注文件的存储路径。
- “annotation-format”：描述标注文件的格式，可选字段，默认为“PASCAL VOC”。目前只支持“PASCAL VOC”。
- “**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 4-12 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 4-13 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4-14**。

表 4-14 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>
```

4.4 从 DLI 导入数据

表格数据集支持从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的详细功能说明，请参考**DLI用户指南**。

图 4-11 DLI 导入数据

- 队列名称：系统自动将当前账号下的DLI队列展现在列表中，用户可以在下拉框中选择需要的队列。
- 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。

说明

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。

4.5 从 MRS 导入数据

ModelArts支持从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。MRS的详细功能说明，请参考**MRS用户指南**。

图 4-12 从 MRS 导入数据

- **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。

4.6 从 DWS 导入数据

ModelArts支持从DWS导入表格数据，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。DWS的详细功能说明，请参考**DWS用户指南**。

图 4-13 从 DWS 导入数据

- **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。

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。

4.7 从本地上传数据

前提条件

- 已存在创建完成的数据集。
- 创建一个空的OBS桶，OBS桶与ModelArts在同一区域，并确保用户具有OBS桶的操作权限。

本地上传

文件型和表格型数据均支持从本地上传。从本地上传的数据存储在OBS目录中，请先提前创建OBS桶。

从本地上传的数据单次最多支持100个文件同时上传，总大小不超过5GB。

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“导入”。

   图 4-14 导入数据

   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”。
   - “数据来源”：“本地上传”
   - “上传数据存储路径”：数据存储的OBS路径。
   - “上传数据”：单击“文件上传”，上传本地的数据，单击“确定”。

图 4-15 从本地上传数据

5 数据分析与预览

用户的原始数据的质量一般无法满足训练的要求，如存在不合法的数据、重复数据等。为了帮助用户提高数据的质量，ModelArts提供了多种能力：

- **数据处理**：提供数据增强、数据清洗、数据校验等能力。
- **自动分组**：通过聚类对数据进行预分类，用户可以根据预分类结果进行标注，有助于均衡不同类别的数据标注数量。
- **数据筛选**：用户可以根据样本属性，自动分组结果等进行数据筛选，帮助用户过滤数据。
- **数据特征分析**：分析数据或者标注结果的特征分布，如图像亮度分布、标注框的分布等，帮助用户分析数据的均衡性，从而提升模型训练的效果。

5.1 数据处理

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。ModelArts提供了数据处理的功能，目的是帮助用户从大量的、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。

ModelArts提供了四种基本的数据处理功能：

- 数据校验：帮助AI开发者提前识别数据中的不合法数据，如已损坏数据、不合格数据等，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：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。
- 数据选择：在AI开发过程中，采集的数据可能存在大量重复数据，重复数据对模型精度提升并没有太大作用，反而需要花费很多时间对其进行标注。使用数据选择进行数据预处理，对采集到的数据去重，根据相似度删除一些重复度比较高的数据。
- 数据增强：数据增强的目的是帮助用户增加数据量。

使用数据处理功能的具体步骤参考**创建数据处理任务**。

5.2 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>数据标注”，进入“数据标注”管理页面。
2. 在标注作业列表中，选择“物体检测”或“图像分类”类型的标注作业，单击标注作业名称进入“标注作业详情页”。
3. 在数据集详情页的“标注>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。
4. 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。

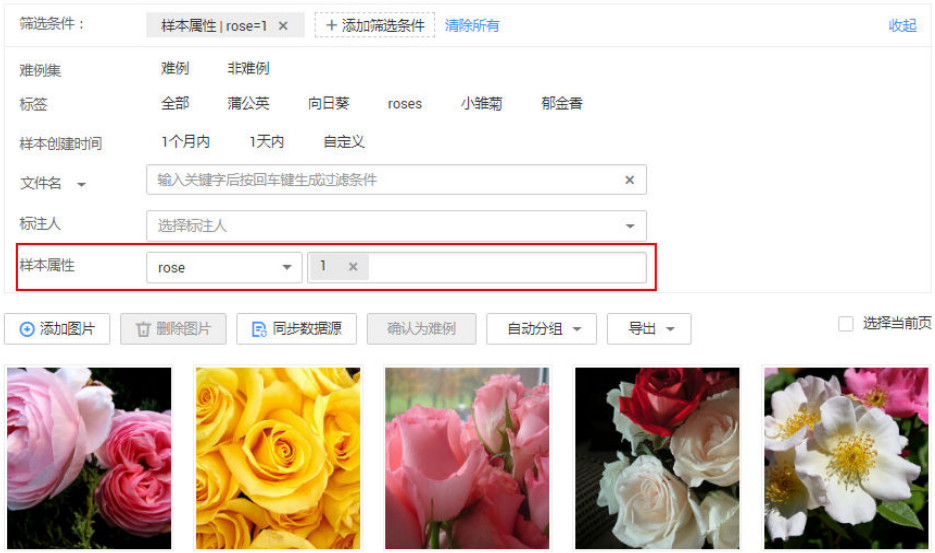
图 5-1 自动分组

5. 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 5-2 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 5-3 自动分组任务历史

5.3 数据筛选

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 5-4 筛选条件

5.4 数据特征分析

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“图片”的数据集，且版本标注类型为“物体检测”和“图像分类”的数据集版本支持数据特征分析。
- 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：
  - 对于标注任务类型为“物体检测”的数据集版本，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。否则，显示已标注的图片的数据特征。
  - 对于标注任务类型为“图像分类”的数据集版本，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。否则，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 5-5 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。

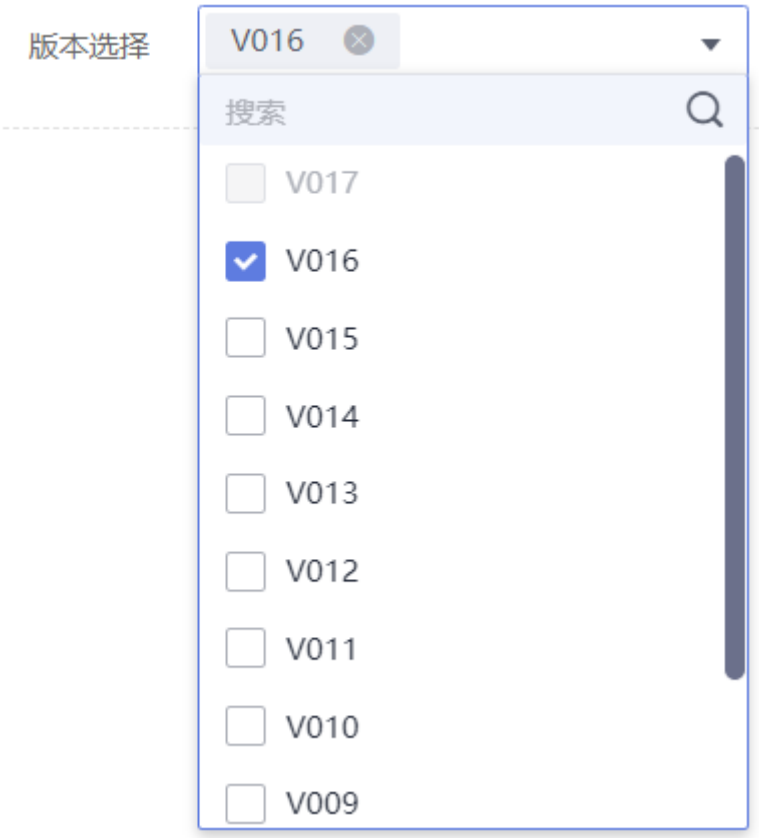
   “版本选择”，即选择当前数据集的已发布版本。

   图 5-6 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 5-7 可选择已执行特征分析的版本



6. 查看数据特征分析结果。

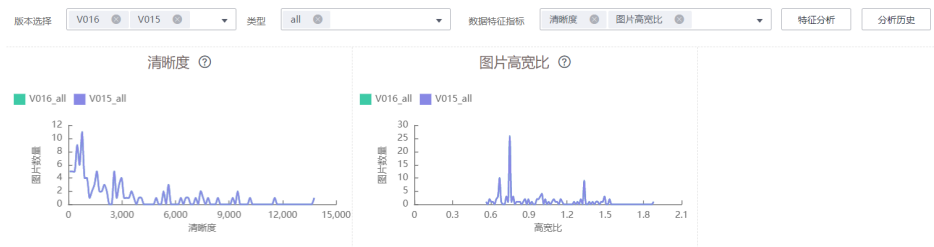
   “版本选择”：在右侧下拉框中选择进行对比的版本。也可以只选择一个版本。

   “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5-8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 5-8 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 5-9 任务历史

支持分析指标及其说明

表 5-1 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。

名称	说明	分析说明
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。（图片总距离表示以图片中心点为起点画一条经过标注框中心点的射线，该射线与图片边界交点到图片中心点的距离）。 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。
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。

6 数据标注

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数据添加标签。您可以通过创建单人标注作业或团队标注作业对数据进行手工标注，或对任务启动智能标注添加标签，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。

- 人工标注：用户创建单人标注作业，对数据进行手工标注。
- 智能标注：在标注一定量的数据情况下，用户可以通过启动智能标注任务对数据进行自动标注，提高标注的效率。
- 团队标注：对于大批量的数据，用户可以通过创建团队标注作业，进行多人协同标注。

关于数据标注的详细信息，请参考**数据标注**。

7 数据发布

7.1 数据发布简介

ModelArts在数据准备过程中，针对同一数据源的数据，对不同时间处理或标注后的数据，按照版本进行区分方便后续模型构建和开发时选择对应的数据集版本进行使用。

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。

7.2 发布数据版本

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“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”。

图 7-1 发布数据集版本

表 7-1 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。

数据集版本文件目录结构

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
```

7.3 管理数据版本

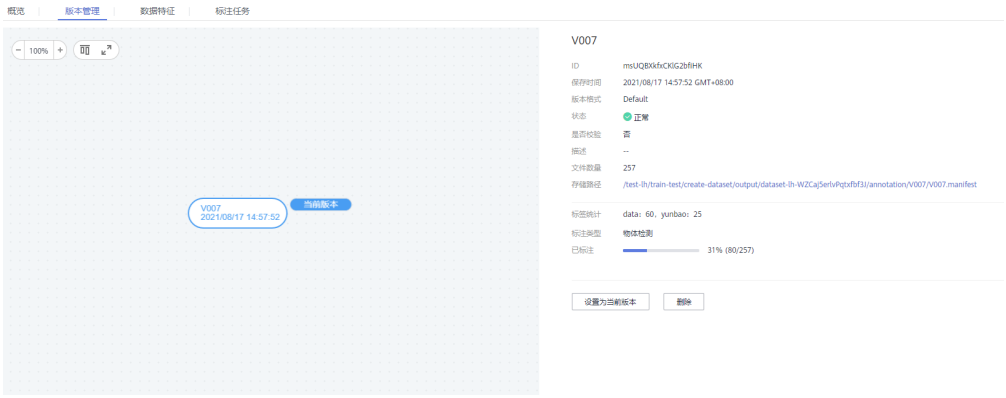
在数据准备的过程中，用户可以将数据发布成多个版本对数据集进行管理。针对已发布生成的数据集版本，用户可以通过查看数据集的演进过程、切换版本、删除版本等操作，对数据集进行管理。

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，并在左侧查看版本演进信息及其发布时间。

图 7-2 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”。

图 7-3 设置当前版本

说明

只有状态为“正常”的版本，才能被设置为当前版本。

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。

8 数据导出

8.1 数据导出简介

针对数据集中的数据，用户可以选中部分数据或者通过条件筛选出需要的数据，导出成新的数据集，或者将数据导出到指定的OBS目录下。用户可以通过任务历史查看数据导出的历史记录。

目前只有“图像分类”、“物体检测”、“图像分割”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。

8.2 导出数据为新数据集

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“图片”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“导出 ”。在弹出的“导出”对话框中，填写相关信息，然后单击“确定”，开始执行导出操作。

   “导出方式”： 选择新数据集。

   “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。

图 8-1 导出新数据集

4. 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。当导出方式选择为新数据集时，在导出成功后，您可以前往“数据集”列表中，查看到新的数据集。
5. 在“数据集概览页”，单击右上角“导出历史”，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。

8.3 导出数据到 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“图片”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“导出 ”。在弹出的“导出”对话框中，填写相关信息，然后单击“确定”，开始执行导出操作。

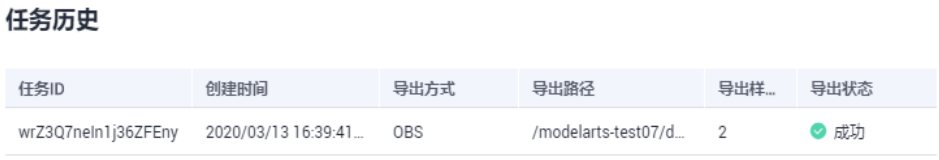
   “导出方式”：选择OBS。

   “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。

图 8-2 导出到 OBS

4. 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。
5. 在“数据集概览页”，单击右上角“导出历史”，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。

图 8-3 任务历史



8.4 导出到 AI Gallery

用户可以将自己的数据发布到AI Gallery，将个人的数据分享给他人使用。用户要发布数据集到AI Gallery，数据集需要有状态为“正常”的数据集版本。

发布数据集到 AI Gallery

1. 选中待发布的数据集，单击“更多”按钮，选择“发布资产”。

2. 在资产发布弹窗中，选择数据集的版本并填写资产发布相关的信息。完成后单击“确定”即可进行发布。

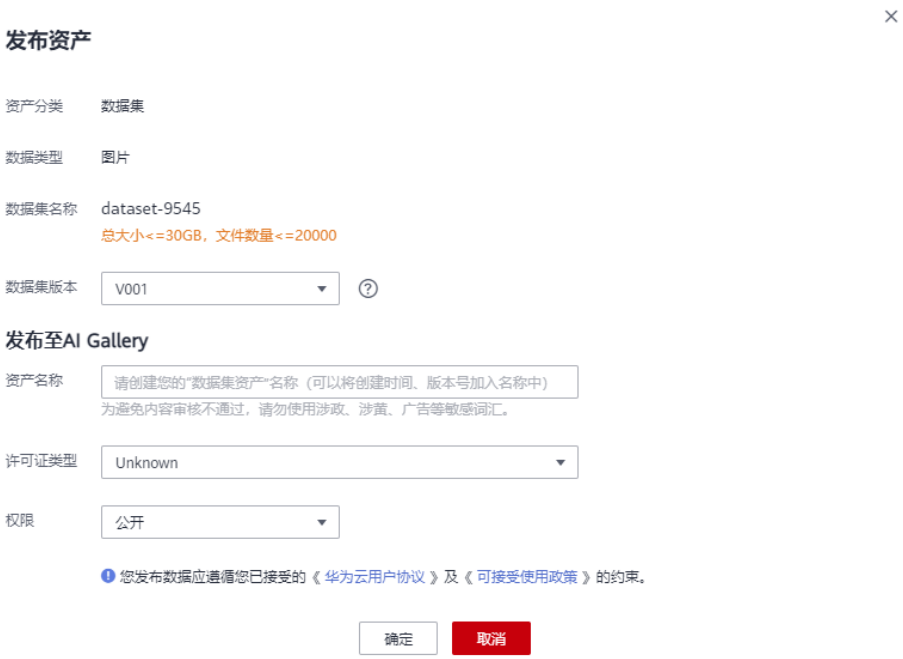
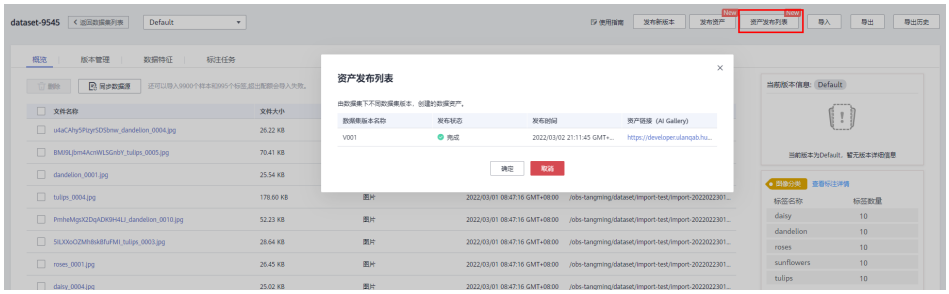


表 8-1 发布数据集到 AI Galley 参数说明

参数	说明
资产标题	在AI Gallery显示的资产名称。
数据集版本	选择目标数据集需要发布的版本。
许可证类型	根据业务需求和数据集类型选择合适的许可证类型。 单击许可证类型后面的感叹号可以查看许可证详情。 **说明** 部分许可证网站说明地址是海外网站，用户可能会因网络限制无法访问。
谁可以看	设置此数据集的公开权限。可选值有： ● “公开”：表示所有使用AI Gallery的用户都可以查看且使用该资产。 ● “指定用户”：表示仅特定用户可以查看及使用该资产。 ● “仅自己可见”：表示只有当前帐号可以查看并使用该资产。

查看数据集资产发布信息

在数据集列表中，单击某个数据集名称进入数据集详情页。选中“资产发布列表”，可以查看资产发布记录。



发布成功后，会生成资产链接，用户可以单击链接跳转到资产的详情页面。

删除发布的数据集

当您需要删除发布在AI Gallery中的数据集时，可以执行如下步骤进行删除。

1. 在AI Gallery页面的右上角单击“个人中心 > 我的数据”。
2. 在“我的发布”页签，单击目标数据集右侧的“删除”，在弹窗中确认删除。

说明

由于数据集是下载至OBS使用的，所以删除已发布的数据集对使用者无影响。