ALIBABA CLOUD

阿里云

机器学习PAI

PAI数据准备

文档版本：20200927

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击设置> 网络> 设置网络类型。
粗体	表示按键、菜单、页面名称等UI元素。	在结果确认页面，单击**确定**。
`Courier字体`	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

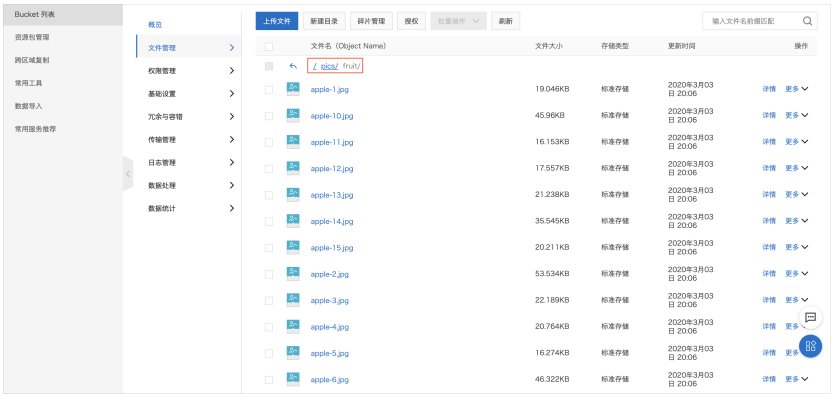
1.注册数据集

PAI支持通过新建数据集和导入数据集文件的方式注册数据集，并通过manifest清单文件管理所有数据集。

通过新建数据集的方式注册数据集

如果原始数据（图像、文本、视频、语音等）存储在OSS，可以通过**新建数据集**的方式注册数据集。系统会遍历指定目录下的同类型文件，并将生成的清单文件存储在指定的OSS目录。

1. 进入**注册数据集**页面。
   i. 登录PAI控制台。
   ii. 在PAI控制台首页，选择**数据准备 > 数据集管理**。
   iii. 在**数据集管理**页面，单击**注册数据集**。
2. 在**注册数据集**页面，配置参数。

参数	描述
数据集名称	长度为1~24个字符，以小写字母、大写字母、数字或中文开头，可以包含下划线（_）或短划线（-）。
注册方式	选择**注册方式**为**新建数据集**。
存储类型	系统默认OSS，不支持修改。
选择存储路径	将选择存储路径配置为一个存储原始数据的OSS文件夹，如下图所示。 
数据类型	系统默认图片，不支持修改。
数据集关键词	便于管理数据集，每个数据集最多添加10个关键词。每个关键词长度为1~10个字符，以小写字母、大写字母、数字或中文开头，可以包含下划线（_）或短划线（-）。

3. 单击**提交**，生成的清单文件内容如下。

```
{"data":{"picUrl":"oss://****/pics/fruit/apple-1.jpg"}}
{"data":{"picUrl":"oss://****/pics/fruit/apple-10.jpg"}}
{"data":{"picUrl":"oss://****/pics/fruit/apple-11.jpg"}}
...
```

通过导入数据集文件的方式注册数据集

如果本地有CSV文件或manifest文件，可以通过导入数据集文件的方式注册数据集。系统将CSV文件转换为manifest文件存储。

1. 进入注册数据集页面。
   i. 登录PAI控制台。
   ii. 在PAI控制台首页，选择数据准备 > 数据集管理。
   iii. 在数据集管理页面，单击注册数据集。
2. 在注册数据集页面，配置参数。

参数	描述
数据集名称	长度为1~24个字符，以小写字母、大写字母、数字或中文开头，可以包含下划线（_）或短划线（-）。
注册方式	选择注册方式为导入数据集文件。
存储类型	系统默认OSS，不支持修改。
选择存储路径	选择OSS的存储目录。
数据类型	将本地CSV或manifest文件拖拽至数据类型下的上传区域。 说明 如果导入的数据集用于标注任务，则数据集的字段名称需要符合标注模板要求，详情请参见图像类。
数据集关键词	便于管理数据集，每个数据集最多添加10个关键词。每个关键词长度为1~10个字符，以小写字母、大写字母、数字或中文开头，可以包含下划线（_）或短划线（-）。

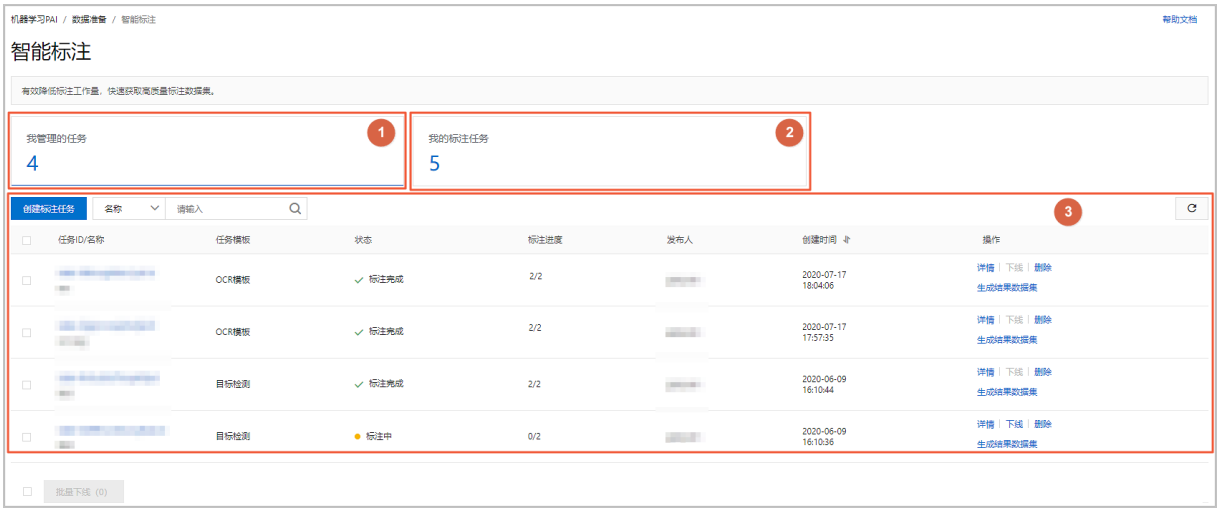
3. 单击提交。

2.智能标注

2.1. 概述

通过智能标注页面，可以查看您以不同角色参与的标注任务。

在智能标注页面，可以查看我管理的任务、我的标注任务及任务列表。



序号	区域	描述
①	我管理的任务	您可以对我管理的任务进行以下操作： • 查看您以管理员身份创建的标注任务。 • 单击任务ID，进入任务详情页面，查看或编辑任务信息。 说明 任务开始标注后，不允许编辑标注策略和标注标签。 • 单击生成结果数据集，系统会在选择的OSS目录下生成标注数据集文件，且每次生成的新数据集文件不会覆盖之前生成的数据集文件。
②	我的标注任务	查看分配给您的标注任务及标注进度。
③	任务列表	显示标注任务的信息，包括任务ID、任务模板、状态、标注进度、发布人、创建时间及操作。

2.2. 图像类

PAI提供了目标检测、语义分割、图像综合标注、OCR及图像分类模板。创建标注任务时，可以根据应用场景选择标注模板。

目标检测

目标检测（Object Detection）任务是对图像中的具体目标进行定位，常用矩形框工具。

- 应用场景

  车辆检测、行人检测及图片搜索等。

- 数据结构
  - 输入数据

    manifest文件的每行数据是一道题目，且每行数据必须包含picUrl字段。

    ```
    {"data":{"picUrl":"oss://****/pics/fruit/apple-1.jpg"}}
    ...
    ```

  - 输出数据

    manifest文件的每行数据由题目和标注结果一起生成。每行数据的JSON结构如下。

    ```
    {
      "data": {
        "picUrl": "oss://****/pics/fruit/apple-1.jpg"
      },
      "label-****(标注任务ID)": {
        "results": [{
          "data": [{
            "id":"Znyumd-*****",
            "type":"image/rectangleLabel",
            "value":{
              "rotation":0,
              "x":40.68320610687023,
              "width":327.52035623409665,
              "y":5.762467474590647,
              "height":296.68117192104745
            },
            "labelColor":"#72bf7d",
            "labels":["apple"]
          }],
          "id":"44****",
          "type":"image"
        }]
      }
    }
    ```

语义分割

语义分割（Semantic Segmentation）任务识别标注图像中存在的内容及位置（通过查找属于它的所有像素）。常用多边形描点工具、笔刷工具及超像素工具。

- 应用场景

  自动驾驶、表情识别及服装分类等。

- 数据结构

- 输入数据

  manifest文件的每行数据是一道题目，且每行数据必须包含picUrl字段。

```
{"data":{"picUrl":"oss://****/pics/fruit/apple-1.jpg"}}
...
```

- 输出数据

  manifest文件的每行数据由题目和标注结果一起生成。每行数据的JSON结构如下。

```
{
  "data": {
    "picUrl": "oss://****/pics/fruit/apple-1.jpg"
  },
  "label-****(标注任务ID)": {
    "results": [{
      "data": [{
        "id":"Znyumd-*****",
        "type":"image/polygonLabel",
        "value":{
          "points": [
            [110, 46],
            [52, 196],
            [48, 168],
            [48, 145],
            [54, 120],
            [63, 93],
            [76, 74]
          ]
        },
        "labelColor":"#72bf7d",
        "labels":["apple"]
      }],
      "id":"44****",
      "type":"image"
    }]
  }
}
```

图像综合标注

图像综合标注（Comprehensive Image Annotation）是指在一组标签集合中，对输入图像的图片内容进行标签匹配。该模板支持使用所有图像标注工具，可以满足更灵活的标注需求。

- 应用场景

  自动驾驶、内容审核及内容识别等。

- 数据结构
  - 输入数据

    manifest文件的每行数据是一道题目，且每行数据必须包含picUrl字段。

    ```
    {"data":{"picUrl":"oss://****/pics/fruit/apple-10.jpg"}}
    ```

  - 输出数据

    manifest文件的每行数据由题目和标注结果一起生成，每行数据的JSON结构如下。

    ```
    {
      "data": {
        "picUrl": "oss://****/pics/fruit/apple-10.jpg"
      },
      "label-****(标注任务ID)": {
        "results": [{
          "data": [{
            "id":"Znyumd-****",
            "type":"image/rectangleLabel",
            "value":{
              "rotation":0,
              "x":40.68320610687023,
              "width":327.52035623409665,
              "y":5.762467474590647,
              "height":296.68117192104745
            },
            "labelColor":"#72bf7d",
            "labels":["红苹果"]
          }],
          "id":"44****",
          "type":"image"
        }]
      }
    }
    ```

OCR模板

光学字符识别OCR（Optical Character Recognition）任务首先将输入图像中的文字转换为文本格式，再根据文字信息类别对输入图像进行分组。

- 应用场景

  证件识别、票据识别、车牌识别及银行卡识别等。

- 数据结构
  - 输入数据

    manifest文件的每行数据是一道题目，且每行数据必须包含picUrl字段。

    ```
    {"data":{"picUrl":"oss://****/img/ocr_card/img0.jpeg"}}
    ```

  - 输出数据

    manifest文件的每行数据由题目和标注结果一起生成，每行数据的JSON结构如下。

    ```
    {
      "data": {
        "picUrl": "oss://****/img/ocr_card/img0.jpeg"
      },
      "label-****(标注任务ID)": {
        "results": [{
          "data": [{
            "direction_of_picture":"downward",
            "type":"ocr/meta"
          },
          {
            "id": "Y4ZFoC-****",
            "direction_of_text": "downward",
            "text": "阿里云计算公司",
            "type": "ocr/polygonLabel",
            "value": {
              "points": [[325.08789110183716,397.47582054138184]]
            },
            "labelColor": "#67bd3a",
            "labels": "公司"
          }],
          "id":"24****",
          "type":"ocr"
        }]
      }
    }
    ```

图像分类

图像分类（Image Classification）是指从一组固定的分类标签集合中，找到与输入图像内容相匹配的一个或多个分类标签，并将其分配给该输入图像。该模板支持单标签和多标签图像分类。

- 应用场景

  相册图片分类、拍照识图、图片搜索及内容推荐等。

- 数据结构
  - 输入数据

    manifest文件的每行数据是一道题目，且每行数据必须包含picUrl字段。

    ```
    {"data":{"picUrl":"oss://****/img/ocr_card/img0.jpeg"}}
    ```

  - 输出数据

    manifest文件的每行数据由题目和标注结果一起生成，每行数据的JSON结构如下。

    ```
    {
      "data": {
        "picUrl": "oss://****/img/ocr_card/img0.jpeg"
      },
      "label-xxxxx(标注任务id)": {
        "results": [{
          "data": [{
            "data":"red",
            "id":"33****",
            "type":"survey/value"
          }],
          "id":"33****",
          "type":"survey"
        }]
      }
    }
    ```

2.3. 标注模板

2.3.1. 视频类

PAI提供了视频分类和物体标记模板。创建标注任务时，可以根据应用场景选择标注模板。

视频分类

视频分类（Video Classification）是指在一组固定的分类标签集合中，找出与输入视频内容相匹配的一个或多个分类标签，并将其分配给该输入视频。该模板支持单标签和多标签分类。

- 应用场景

  视频监控、直播推荐及短视频推荐等。

- 数据结构

- 输入数据

  manifest文件的每行数据是一道题目，且每行数据必须包含videoUrl字段。

  ```
  {"data":{"videoUrl":"oss://xxxxxxxxx.oss-cn-shanghai-internal.aliyuncs.com/video/English.mp4"}}
  ```

- 输出数据

  manifest文件的每行数据由题目和标注结果一起生成。每行数据的JSON结构如下。

  ```
  {
   "data": {
    "videoUrl": "oss://xxxxxxxxxxx.oss-cn-shanghai-internal.aliyuncs.com/video/show.webm"
   },
   "label-bo0pc51ok7olwe****": {
    "results": [{
     "data": [{
      "data": ["汽车推广", "宣传片"],
      "id": 165****,
      "type": "survey/multivalue"
     }],
     "id": "165****",
     "type": "survey"
    }]
   }
  }
  ```

物体标记

物体标记（Object Marking）是指对视频中某一帧或某些帧出现的特定物体进行定位并标记，常用矩形框工具或多边形框工具。

- 应用场景

  自动驾驶、安防监控及视频推荐等。

- 数据结构
  - 输入数据

    manifest文件的每行数据是一道题目，且每行数据必须包含videoUrl字段。

    ```
    {"data":{"videoUrl":"oss://xxxxxxxxx.oss-cn-shanghai-internal.aliyuncs.com/video/English.mp4"}}
    ```

  - 输出数据

    manifest文件的每行数据由题目和标注结果一起生成，标注结果中包含多帧（Frame）的标注结果。每行数据的JSON结构如下。

    ```
    {
     "data": {
      "videoUrl": "oss://xxxxxxxxx.oss-cn-shanghai-internal.aliyuncs.com/video/English.mp4"
    ```

```
  },
  "label-rv0ih5l409u9x9****": {
   "results": [{
    "data": [{
     "speed_of_play": 1,
     "duration": 300.733375,
     "width": 960,
     "number_of_frames": 9022,
     "type": "video/meta",
     "frame_rate": 30,
     "height": 540
    }, {
     "frames": {
      "frame-443": {
       "L6G-5Oc5STCSOuzKb****": {
        "rotation": 0,
        "x": 336.0483870967742,
        "width": 488.4677419354839,
        "y": 108.87096774193549,
        "height": 349.83870967741933
       }
      }
     },
     "type": "video/frames"
    }, {
     "frames": [443],
     "custom": {
      "这是什么类型的视频": "学习英语视频",
      "视频名称": "英语学习",
      "视频类别": ["教育片"]
     },
     "id": "L6G-5Oc5STCSOuzKb****",
     "type": "video/rectangleLabel",
     "value": "",
     "labelColor": "#ff7700"
    }],
    "id": "129****",
    "type": "video"
   }]
  }
}
```

```
}
```

2.3.2. 文本类

PAI提供了文本分类标注模板，本文为您介绍其应用场景及数据结构。

文本分类

文本分类（Text Classification）是指在一组固定的分类标签集合中，找到与输入文本内容相匹配的一个或多个分类标签，并将其分配给该输入文本。该分类模板支持单标签和多标签。

- 应用场景

  新闻推荐、知识管理及垃圾信息过滤等。

- 数据结构
  - 输入数据

    manifest文件的每行数据是一道题目，且每行数据必须包含content字段。

    ```
    {"data":{"content":"欢迎使用机器学习PAI！"}}
    ```

  - 输出数据

    manifest文件的每行数据由题目和标注结果一起生成。每行数据的JSON结构如下。

    ```
    {
     "data": {
      "content": "欢迎使用机器学习PAI！"
     },
     "label-p9fypszvupm9bh****": {
      "results": [{
       "data": [{
        "data": "积极内容",
        "id": 112****,
        "type": "survey/value"
       }],
       "id": "112****",
       "type": "survey"
      }]
     }
    }
    ```

2.4. 创建标注任务

本文为您介绍如何创建标注任务。

前提条件

注册数据集，详情请参见注册数据集。

操作步骤

1. 登录PAI控制台。
2. 在PAI控制台首页，选择数据准备 > 智能标注。
3. 在智能标注页面，单击创建标注任务。
4. 在选择模板配置向导页面，配置参数，并单击下一步。

参数		描述
选择模板		选择标注任务的模板，系统支持的模板包括： ○ 图像类 ■ 目标检测 ■ 语义分割 ■ 图像综合标注 ■ OCR模板 ■ 图像分类 ○ 文本类 支持文本分类模板。 ○ 视频类 ■ 视频分类 ■ 物体标记
分类选项		图片分类标签。当选择模板为目标检测、语义分割或图像综合标注时，该参数生效。
标注整图方向	是否需要标注整图方向	标注整图方向的开关。仅当选择模板为OCR模板时，该参数生效。
	标注时可选文字方向	标注文字方向的开关。如果图片中的文字方向均与整图方向一致，则关闭该开关，从而简化标注步骤。仅当选择模板为OCR模板时，该参数生效。
文字类别		文本信息类别。仅当选择模板为OCR模板时，该参数生效。
添加自定义标签		文本额外的类别标签。仅当选择模板为OCR模板时，该参数生效。
标注类型		支持单标签和多标签图像分类，结合实际标注需求进行选择。当选择模板为图像分类时，该参数生效。
标注标签		分类标签，系统会使用不同颜色显示每个标签。当选择模板为图像分类时，该参数生效。

5. 在基础信息配置向导页面，配置参数，并单击下一步。

参数	描述
任务名称	长度为1~30个字符，以小写字母、大写字母、数字或中文开头，可以包含下划线（_）或短划线（-）。

参数	描述
任务描述	长度为1~64个字符，以小写字母、大写字母、数字或中文开头，可以包含下划线（_）或短划线（-）。
输入数据集	可以选择多个数据集组合创建标注任务，所选数据集需要与标注主题相关。如果无可用数据集，则单击**输入数据集**后的**注册数据集**，进行数据集创建。
输出数据集位置	存储标注结果的OSS路径。标注任务过程中，每次单击**生成结果数据集**，都会在该路径生成一份截止当前的标注结果数据集。

6. 在**标注策略配置**向导页面，配置参数，并单击**提交**。

参数	描述
发题策略	系统默认**标注员按次领取固定数量**，不支持修改。
每次领取	标注员每次领取的题目数量。 说明 每次领取的题目数量可以小于总题目数除以总人数，让标注效率高的标注员可以标注到更多题目，提高总体标注效率。
添加标注员	可以添加多个标注员，系统支持主账号及其所有子账号协作完成标注任务。

2.5. 标注图像

本文为您介绍如何标注图像。

前提条件

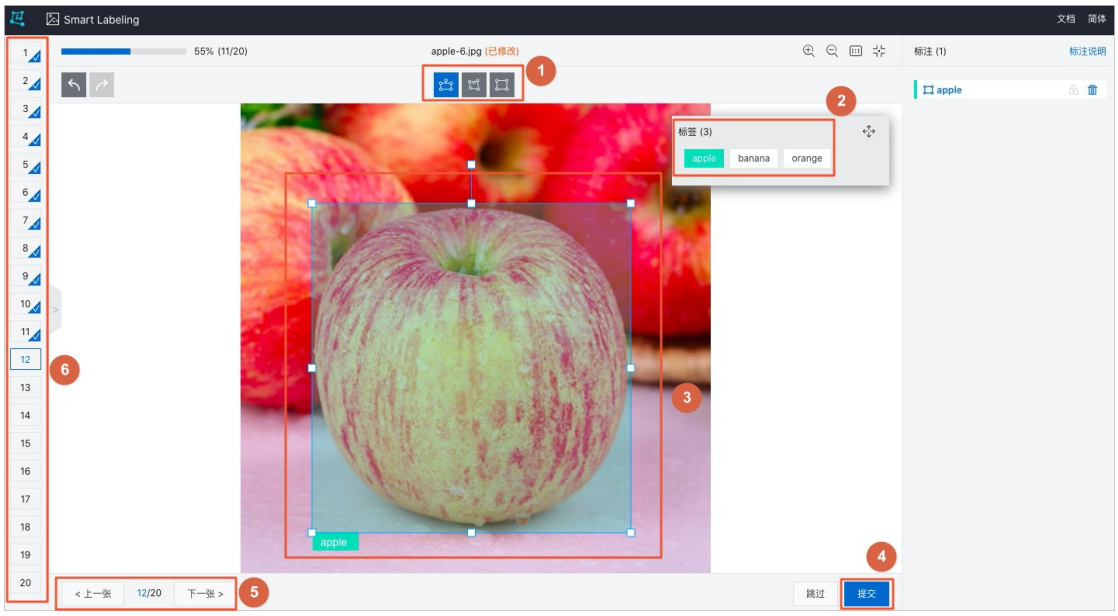
创建标注任务或管理员分配标注任务，详情请参见创建标注任务。

操作步骤

1. 进入标注页面。
   i. 登录PAI控制台。
   ii. 在PAI控制台首页，选择**数据准备 > 智能标注**。
   iii. 在**智能标注**页面，单击**我的标注任务**。
   iv. 在任务列表，单击待标注任务**操作**列下的**去标注**。
2. 标注图像。
   i. 在标注页面，单击图标，选择标注工具。
   ii. 在**标签**区域，选中一个标签。

   说明 选中标签后，如果您没有手动切换标签，则以相同的标签进行标注。

iii. 在图像中，使用标注工具进行标注。



iv. （可选）如果无可标注内容，单击**跳过**。

v. 单击**提交**。

vi. 您可以通过以下任何一种方式，浏览题目，并标注所有题目：

- 在标注页面下方，单击**上一张**或**下一张**。
- 在标注页面左侧，单击任务列表缩略图。

2.6. 标注快捷键

PAI支持使用快捷键标注图像，提高操作效率。

功能	快捷键
选择矩形工具	R
选择多边形工具	P
切换标签	1、2、3、4、5、6、7、8、9（仅支持前9个标签）
跳过当前题目	S
提交当前题目	Enter
上一张	Left
下一张	Right

> **说明** 将鼠标悬停至功能按钮，可以查看其对应的快捷键。

2.7. 智能贴合与智能识别

传统的图像文本标注工具需要手动精确贴合文本，不仅标注效率低，而且在文字密集的情况下容易造成误标注。PAI智能标注支持文本框智能贴合，可以自动校正倾斜、变形或密集的文本框。同时，PAI智能标注支持智能识别，可以对框选的内容进行自动识别并生成对应文字，帮助您快速完成标注任务。

文本智能贴合

- 传统的手动标注

  传统的手动标注难以控制标注框大小，标注效率低。

- 智能标注

  智能标注可以使文本框自动收缩至紧贴文字，提高了标注速度及质量。

文本智能识别

- 传统的手动标注

框选文本内容后，需要手动输入对应的文字，标注效率低。

- 智能标注

  框选文本内容后，单击**智能识别**，系统可以自动识别文本并生成框选区域对应的文字。

倾斜文本标注

- 传统的手动标注

需要使用多边形工具进行标注，标注效率低。

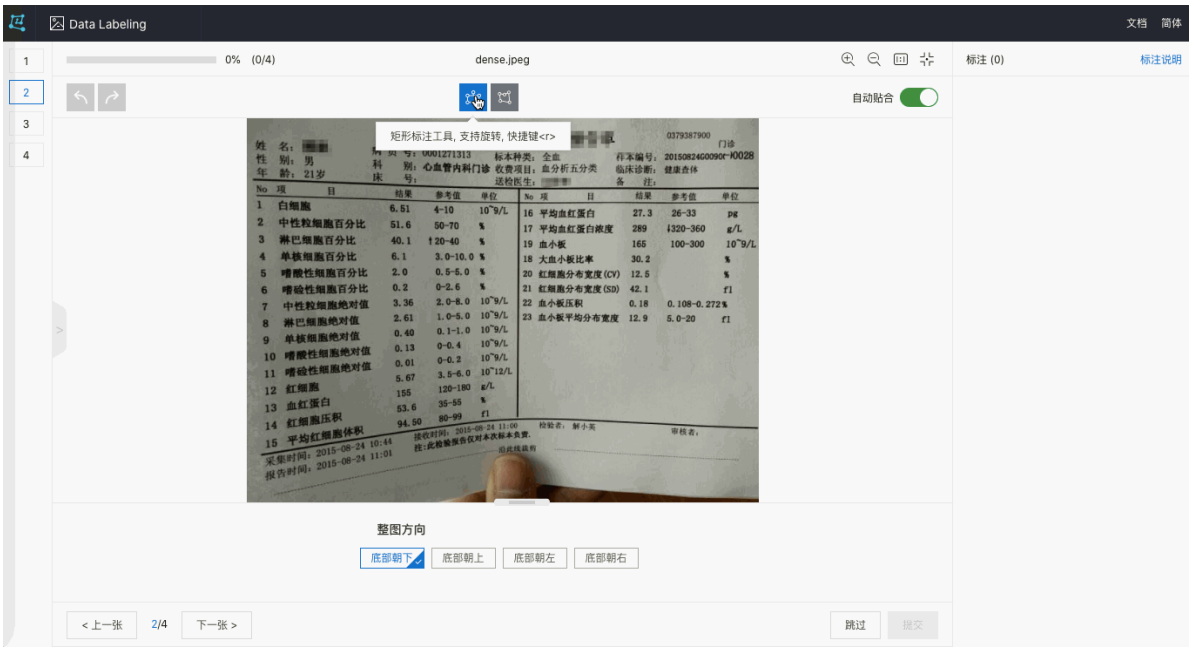
- 智能标注

  只需要使用矩形工具标注倾斜文本区域，系统可以自动识别文本范围并自动贴合文本。

密集文本标注

- 传统的手动标注

通过传统手动标注的方式进行密集文本标注，通常会框选到目标文本周围的文字，易造成误标注。



- 智能标注

通过智能标注的方式进行密集文本标注，只需要使用矩形工具大范围地框选密集文本区域。即使误框选了其他文本，系统也可以自动贴合目标文本。