

Fusion AI Space

24.0.1

# 操作指南

文档版本 01

发布日期 2024-06-30

**版权所有 © 超聚变数字技术有限公司 2024。保留一切权利。**

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

### **商标声明**

**XFUSION** 和其他超聚变商标均为超聚变数字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**

本文中，只是为了描述的简洁和方便理解，用“xFusion”指代“xFusion Digital Technologies Co., Ltd.”，这并不代表“xFusion”还可以具备其它含义。基于本文中单独提及或描述的“xFusion”，不能用于“xFusion Digital Technologies Co., Ltd.”之外的理解或表达，超聚变数字技术有限公司也不承担因单独使用“xFusion”所带来的其它任何法律责任。

您购买的产品、服务或特性等应受超聚变数字技术有限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，超聚变数字技术有限公司对本文档内容不做任何明示或默示的声明或保证。

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

## **超聚变数字技术有限公司**

地址：河南省郑州市郑东新区龙子湖智慧岛正商博雅广场1号楼9层 邮编：450046

网址：<https://www.xfusion.com>

# 前言

## 概述

本文档介绍了AI Space的操作方法，用于指导用户（包括组织管理员、部门管理员和普通用户）使用AI平台开展AI业务。

## 读者对象

本文档主要适用于以下工程师：

- 技术支持工程师
- 企业管理员

## 符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

| 符号 | 说明  |
|----|---|
|    | 表示如不避免则将会导致死亡或严重伤害的具有高等级风险的危害。  |
|    | 表示如不避免则可能导致死亡或严重伤害的具有中等级风险的危害。  |
|    | 表示如不避免则可能导致轻微或中度伤害的具有低等级风险的危害。  |
|    | 用于传递设备或环境安全警示信息。如不避免则可能会导致设备损坏、数据丢失、设备性能降低或其它不可预知的结果。<br>“须知”不涉及人身伤害。 |
|    | 对正文中重点信息的补充说明。<br>“说明”不是安全警示信息，不涉及人身、设备及环境伤害信息。                       |

## 修改记录

| 文档版本 | 发布日期       | 修改说明     |
|------|------------|----------|
| 01   | 2024-06-30 | 第一次正式发布。 |

# 目 录

|                 |    |
|-----------------|----|
| <b>前言</b>       | ii |
| <b>1 产品介绍</b>   | 1  |
| <b>2 新手入门</b>   | 2  |
| 2.1 用户登录        | 2  |
| 2.2 界面说明        | 4  |
| 2.3 初始设置        | 5  |
| <b>3 操作指南</b>   | 11 |
| 3.1 AI 平台       | 11 |
| 3.1.1 业务流程      | 11 |
| 3.1.2 概览        | 12 |
| 3.1.3 数据管理      | 13 |
| 3.1.3.1 文件管理    | 13 |
| 3.1.3.2 数据集管理   | 18 |
| 3.1.3.2.1 准备数据  | 21 |
| 1. 视觉类数据        | 21 |
| 2. 文本类数据        | 24 |
| 3. 音频类数据        | 25 |
| 3.1.3.2.2 创建数据集 | 27 |
| 3.1.3.2.3 数据标注  | 35 |
| 1. 自动标注         | 35 |
| 2. 手动标注         | 37 |
| 1. 图像分类         | 37 |
| 2. 目标检测         | 38 |
| 3. 语义分割         | 39 |
| 4. 目标跟踪         | 40 |
| 5. 文本标注         | 41 |
| 6. 音频标注         | 45 |
| 3.1.3.2.4 数据增强  | 45 |
| 3.1.3.2.5 发布数据集 | 50 |
| 3.1.3.2.6 预置数据集 | 53 |
| 3.1.3.3 标签组管理   | 53 |
| 3.1.3.3.1 我的标签组 | 54 |

|                                |     |
|--------------------------------|-----|
| 3.1.3.3.2 预置标签组.....           | 60  |
| 3.1.3.4 标注服务管理.....            | 60  |
| 3.1.4 算法开发.....                | 69  |
| 3.1.4.1 算法开发简介.....            | 69  |
| 3.1.4.2 Notebook.....          | 69  |
| 3.1.4.2.1 Notebook 镜像打包指导..... | 77  |
| 1. Nvidia 镜像与普通镜像处理步骤.....     | 77  |
| 2. 昇腾镜像与非 root 用户镜像处理.....     | 82  |
| 3.1.4.3 算法管理.....              | 87  |
| 3.1.5 训练管理.....                | 97  |
| 3.1.5.1 训练管理简介.....            | 97  |
| 3.1.5.2 训练任务.....              | 98  |
| 3.1.5.2.1 任务模板.....            | 106 |
| 3.1.5.3 可视化任务.....             | 108 |
| 3.1.6 模型管理.....                | 109 |
| 3.1.6.1 模型管理简介.....            | 109 |
| 3.1.6.2 模型列表.....              | 110 |
| 3.1.6.2.1 我的模型.....            | 110 |
| 3.1.6.2.2 预训练模型.....           | 117 |
| 3.1.7 推理服务.....                | 119 |
| 3.1.7.1 推理服务简介.....            | 119 |
| 3.1.7.2 推理镜像预处理（平台框架）.....     | 121 |
| 3.1.7.3 在线服务.....              | 122 |
| 3.1.7.3.1 查看在线服务列表.....        | 122 |
| 3.1.7.3.2 创建在线服务.....          | 124 |
| 3.1.7.3.3 启停在线服务.....          | 131 |
| 3.1.7.3.4 预测在线服务.....          | 133 |
| 3.1.7.3.5 编辑在线服务.....          | 135 |
| 3.1.7.3.6 回滚在线服务.....          | 136 |
| 3.1.7.3.7 删除在线服务.....          | 137 |
| 3.1.7.4 批量服务.....              | 138 |
| 3.1.7.4.1 查看批量服务列表.....        | 138 |
| 3.1.7.4.2 创建批量服务.....          | 139 |
| 3.1.7.4.3 启停批量服务.....          | 144 |
| 3.1.7.4.4 编辑批量服务.....          | 145 |
| 3.1.7.4.5 Fork 批量服务.....       | 146 |
| 3.1.7.4.6 删除批量服务.....          | 147 |
| 3.1.7.4.7 下载批量服务结果.....        | 148 |
| 3.1.8 镜像管理.....                | 149 |
| 3.1.8.1 查看镜像.....              | 149 |
| 3.1.8.2 创建镜像.....              | 151 |
| 3.1.8.2.1 录入镜像.....            | 151 |

|                               |     |
|-------------------------------|-----|
| 3.1.8.2.2 上传镜像.....           | 154 |
| 3.1.8.2.3 制作镜像.....           | 156 |
| 3.1.8.3 编辑镜像.....             | 158 |
| 3.1.8.4 设置 Notebook 默认镜像..... | 160 |
| 3.1.8.5 下载镜像.....             | 161 |
| 3.1.8.6 转为公共镜像.....           | 163 |
| 3.1.8.7 分享镜像.....             | 163 |
| 3.1.8.8 删除镜像.....             | 164 |
| 3.1.9 控制台.....                | 165 |
| 3.1.9.1 队列管理.....             | 166 |
| 3.1.9.2 资源规格管理.....           | 170 |
| 3.1.9.2.1 基础资源.....           | 170 |
| 3.1.9.2.2 GPU.....            | 173 |
| 1. 整卡.....                    | 173 |
| 2. MIG.....                   | 176 |
| 3.1.9.2.3 NPU.....            | 179 |
| 1. 整卡.....                    | 179 |
| 2. vNPU.....                  | 181 |
| 3.2 运营.....                   | 183 |
| 3.2.1 资源定价.....               | 183 |
| 3.2.2 财务管理.....               | 183 |
| 3.2.2.1 收支明细.....             | 184 |
| 3.2.2.3 集群报表.....             | 185 |
| 3.2.2.3.1 作业详情报表.....         | 185 |
| 3.2.2.3.2 完成作业报表.....         | 187 |
| 3.2.4 账单管理.....               | 190 |
| 3.2.4.1 我的账单.....             | 190 |
| 3.2.4.2 所有账单.....             | 192 |
| 3.3 运维.....                   | 194 |
| 3.3.1 日志.....                 | 194 |
| 3.3.1.1 审计日志.....             | 194 |
| 3.4 系统设置.....                 | 196 |
| 3.4.1 用户.....                 | 196 |
| 3.4.1.1 用户信息.....             | 196 |
| 3.4.1.2 用户管理.....             | 196 |
| 3.4.1.3 成员管理.....             | 203 |
| 3.4.1.4 角色管理.....             | 205 |
| 3.4.1.4.1 普通模式.....           | 205 |
| 3.4.1.4.2 三员模式.....           | 207 |
| 3.4.2 安全.....                 | 210 |
| 3.4.2.1 双因素认证.....            | 210 |
| 3.4.3 节点组.....                | 211 |

|  |            |
|--|------------|
| 3.4.4 许可证.....   | 212        |
| <b>4 最佳实践.....</b>   | <b>215</b> |
| <b>5 常见问题.....</b>   | <b>216</b> |
| 5.1 GPU 监控为零、GPU 大屏无数据.....  | 216        |
| 5.2 创建节点标签.....  | 218        |
| 5.3 进行下载操作时没有反应，下载失败.....  | 218        |
| 5.4 创建 AI 作业，提示 “Read timed out executing POST http://admin/billing/created” ..... | 219        |
| 5.5 查看正在运行中的训练任务日志，显示 “暂无日志” .....   | 220        |
| 5.6 获取组织的 PVC Volume.....  | 221        |
| <b>6 术语&amp;缩略语.....</b>   | <b>223</b> |
| <b>7 如何获取帮助.....</b>   | <b>224</b> |
| 7.1 收集必要的故障信息.....   | 224        |
| 7.2 如何使用文档.....  | 224        |
| 7.3 获取技术支持.....  | 224        |

# 1 产品介绍

AI Space是面向AI模型生产的生命周期，推出的集数据管理（数据集管理、智能标注和数据增强）、模型开发、模型训练和模型管理等功能于一体的AI算法构建平台，旨在降低用户算法开发门槛，快速推出面向各行各业的智能算法。

## 多租户角色介绍

AI Space支持多租户体系管理，目前支持的角色包括：超级管理员、组织管理员、部门管理员、普通用户，不同角色对应不同权限。

- 超级管理员（Administrator）：超级管理员是整个多租户系统的最高权限角色，负责管理整个系统服务器配置、安全性、租户的管理与配额设置、系统管理角色。负责系统级别的设置，如升级、系统配置、集群管理、许可证、监控日志等。
- 组织管理员（TenantManageRole）：组织管理员是各个组织的管理者，他们对所属组织内部资源和成员具有较高的权限。组织管理员可以创建和管理组织内的用户、部门管理、创建租户内角色、配额设置等。他们是组织内部的关键联系人，负责与超级管理员和部门管理员进行协调和沟通。
- 部门管理员（DepartmentManageRole）：部门管理员是租户内不同部门或团队的管理者，负责管理属于自己部门的资源和成员。部门管理员管理本部门的用户账号、配额管理，并确保部门内部的数据安全和操作合规性。他们通常与组织管理员合作，确保各个部门的需求得到满足，并在必要时与超级管理员进行沟通。
- 普通用户（OrdinaryUserRole）：普通用户是租户内的普通成员，他们使用多租户系统的功能和资源来完成自己的工作任务。普通用户通常具有最低级别的权限，只能访问和操作与自己工作相关的资源，主要是AI相关的业务，包括数据管理、算法开发、训练管理、推理部署等，但对于系统的整体配置和管理是无权访问的。

# 2 新手入门

[2.1 用户登录](#)

[2.2 界面说明](#)

[2.3 初始设置](#)

## 2.1 用户登录

### 前提条件

- 登录AI Space建议使用Google Chrome 80.0及以上版本的浏览器，不建议使用Internet Explorer浏览器。
- 已配置维护终端与本系统的网络相通。
- 已获取系统WebUI的访问地址、用户名、密码。
- 如果使用LDAP域用户登录，请确保LDAP服务器与系统之间通信正常，且已在系统上启用LDAP功能并配置好LDAP服务器和用户组信息。LDAP配置请参见《AI Space 管理员指南》LDAP配置章节。
- 如果使用DNS域名登录，请确保DNS服务器与系统间通信正常，且已在系统上配置好域名和DNS服务器相关信息。DNS配置请参见《AI Space 管理员指南》DNS配置章节。

### 注意事项

- AI Space WebUI最多支持100路会话同时进行。
- 默认情况下，系统会话超时时间为10分钟，即在10分钟内，如果您未在WebUI执行任何操作，系统将自动登出，此时需输入用户名和密码重新登录。
- 30秒内连续3次登录某个帐户失败，系统将对当前IP地址下该帐户进行锁定，默认锁定5分钟，若30秒内导致3个帐户锁定，系统将对该IP地址进行锁定，默认锁定15分钟。
- 为保证系统的安全性，首次登录时可以修改初始密码，并定期更新密码。
- 对于“Administrator”、“SysAdmin”、“SecAdmin”、“AuditAdmin”和“rootRedfish”用户，系统不支持密码找回功能。忘记密码后，需要重新安装。因此，请务必牢记“Administrator”、“SysAdmin”、“SecAdmin”、“AuditAdmin”和“rootRedfish”用户的密码。

## 操作步骤

**步骤1** 在PC端打开浏览器，在地址栏中输入“<https://系统WebUI的访问地址>”并按“Enter”键。

### 说明

- 系统WebUI的访问地址可以是如下形式：
  - IPv4地址：格式为XXX.XXX.XXX.XXX。
  - IPv6地址：格式为[XXXX:XXXX:XXXX:XXXX:XXXX:XXXX:XXXX:XXXX]。
  - DNS域名：系统的DNS域名全称。
- 输入IPv6地址时，必须使用[ ]将其括起来，而IPv4地址无此限制。
- 浏览器可能会提示网站的安全证书有问题，此时只要确认系统WebUI的访问IP地址正确，您仍然可以选择忽略该提示并继续访问。

**步骤2** 输入登录信息。

在登录界面中需要输入的信息如**表2-1**所示。

**表 2-1 登录参数**

| 参数  | 说明   |
|-----|--|
| 用户名 | 登录用户的用户名。支持的用户名包括： <ul style="list-style-type: none"><li>本地用户：支持输入长度为6~32个字符的用户名。</li><li>LDAP用户：支持输入最大长度为255个字符的用户名。</li><li>NIS用户：支持输入最大长度为512个字符的用户名。</li></ul> |
| 密码  | 登录用户的密码。为了保证安全，用户应定期修改自己的登录密码。   |
| 域名  | <ul style="list-style-type: none"><li>当使用本地用户登录时，可选择“本地”域名。</li><li>当使用LDAP用户登录时，可直接选择LDAP域名。</li><li>当使用NIS用户登录时，可直接选择NIS域名。</li></ul>                            |

### 说明

- 普通模式下登录AI Space WebUI的默认用户名为“Administrator”，默认密码为“Admin@9000”。
- 三员模式下登录AI Space WebUI的默认用户名为“SysAdmin”或“SecAdmin”或“AuditAdmin”，默认密码为“Admin@9000”。
- 调用REST API的默认用户名为“rootRedfish”，默认密码为“Machine@123”。
- 当用户名或密码输入错误，再次输入时，需要输入验证码。若验证码不清晰，可单击刷新。登录验证码有效期为30秒，超时后需要刷新验证码重新输入。
- 当密码连续3次输入错误时，帐号会被自动锁定5分钟，锁定时间满5分钟后自动解锁。

**步骤3** 单击“登录”。

### 📖 说明

- 非首次登录：在执行此步骤后，直接进入系统首页。
- 首次登录：在执行此步骤后，请继续执行步骤**步骤4至7**，进入系统首页。

**步骤4** 按照提示信息输入初始密码、新密码以及确认密码，单击“确认”。

### 📖 说明

密码复杂度检查规则如下：

- 长度为8~32个字符。
- 至少包含一个空格或者以下特殊字符：  
`~!@#\$%^&\*(\*)\_+=|[{}];:"<,>/?
- 至少包含以下字符中的两种：
  - 小写字母：a ~ z
  - 大写字母：A ~ Z
  - 数字：0 ~ 9
- 不能包含用户名或用户名倒写。
- 新旧密码至少在2个字符位上不同。
- 新密码不能与前N个历史密码相同。N的取值可以参见《AI Space 管理员指南》安全策略章节获取。
- 不能为弱密码字典中的密码。弱密码字典请参见《AI Space 管理员指南》安全策略章节。

**步骤5** 在登录界面，重新输入用户名和修改后的密码，单击“登录”。

**步骤6** 单击“登录”，弹出隐私声明界面。

**步骤7** 单击“同意”，进入系统首页。

----结束

## 2.2 界面说明

AI Space界面由下图所示，各组成区域作用如**表2-2**所示。

图 2-1 AI Space 平台界面

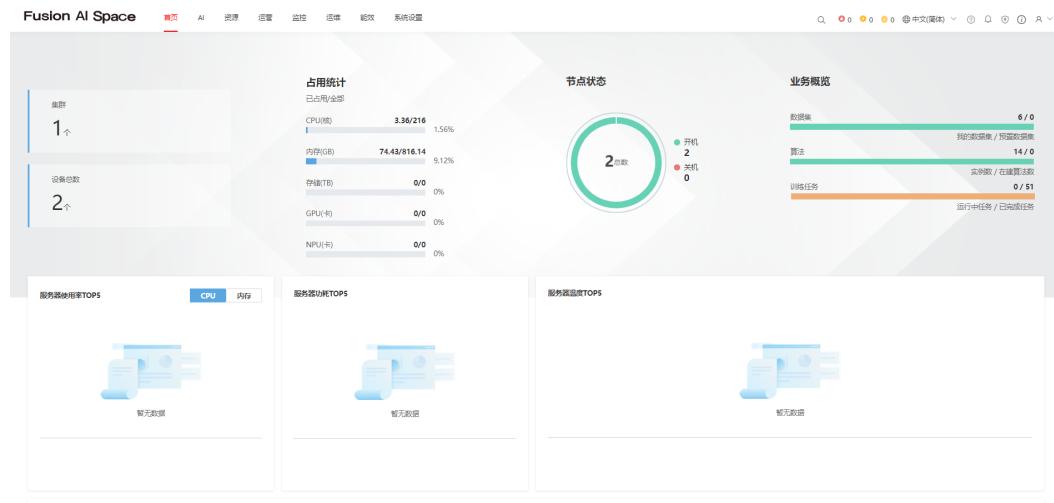


表 2-2 AI Space 界面组成

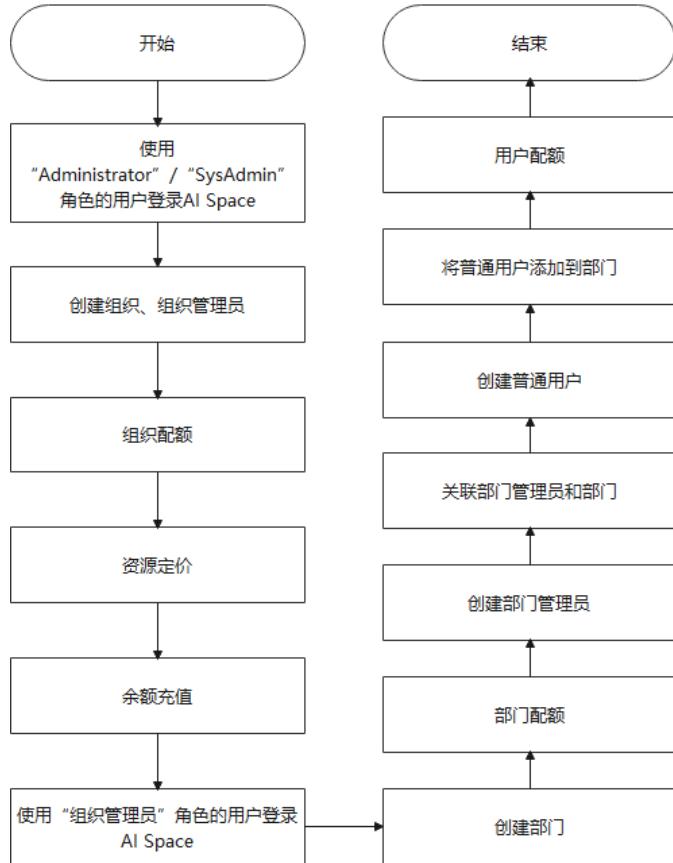
| 名称   | 说明  |
|------|---|
| 当前用户 | 展示当前登录用户和登录时间，并提供查看基本信息、修改密码的操作入口。单击  ，可以退出界面。 |
| 版本信息 | 展示当前版本和版权信息。  |
| 隐私声明 | 超聚变隐私声明。  |
| 任务   | 展示当前正在执行的任务，提供任务中心的快捷入口。  |
| 帮助   | 联机帮助入口，提供界面信息和操作的在线帮助。  |
| 语言   | 界面显示的语言种类，支持切换中文（简体）和英文。  |
| 告警   | 展示当前管理的所有设备的各个级别告警数量。   |
| 搜索   | 可全局搜索当前AI Space中的所有内容，且搜索内容不超过64个字符。  |
| 导航栏  | 提供首页、AI、资源、运营、监控、运维、能效和系统设置的操作入口。如果未授权您查看或操作某个入口，则不会在导航栏中显示。  |

## 2.3 初始设置

本章节指导用户首次登录AI Space后，在开展AI业务前需要进行的初始设置，包括创建组织、组织配额等。

## 初始设置流程

图 2-2 初始设置流程图



## 操作步骤

**步骤1** 参考[2.1 用户登录](#)，普通模式下使用“Administrator”角色用户，三员模式下使用“SysAdmin”角色用户登录AI Space界面。

**步骤2** 创建组织、组织管理员。

- 依次单击“系统设置 > 用户 > 组织管理”，进入组织管理界面。
- 单击“创建组织”。  
弹出创建组织窗口。
- 设置组织信息参数。

表 2-3 组织信息参数说明

| 参数   | 说明   |
|------|--|
| 基本信息 |  |
| 组织名称 | 组织的名称。<br>长度为4~32个字符，仅支持小写字母、数字和特殊字符“-”，开头和结尾不支持特殊字符“-”。 |

| 参数    | 说明   |
|-------|--|
| 组织标识  | 组织的标识。<br>长度为4~32个字符，仅支持小写字母、数字和特殊字符“-”，开头和结尾不支持特殊字符“-”。 |
| 描述    | 组织的描述信息。   |
| 启用状态  | 是否启用该组织。   |
| 管理员信息 |  |
| 用户名   | 组织管理员用户的用户名。用户名只能包含大小写字母或数字，且长度为6~32位。                   |
| 密码    | 组织管理员用户的密码。  |
| 密码确认  | 再次输入设置的密码。   |

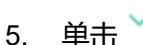
4. 单击“确定”。

#### 步骤3 组织配额。

- 在目标组织所在行，单击“配额”。
- 进入配额界面，在配额界面可以查看资源配额详情。

表 2-4 资源配额详情

| 参数          | 说明  |
|-------------|---|
| 资源类型        | 资源的类型，包括基础资源、GPU、MIG、NPU、vNPU。  |
| 资源名称        | 资源的名称。  |
| 已用配额        | 显示当前已使用资源数量及百分比。  |
| 可用配额        | 可使用配额数量。<br><b>说明</b><br>请确保组织的基础资源（CPU/内存/存储/镜像配额）可用配额大于0，否则组织下用户无法正常使用AI功能。 |
| 组织队列资源可预留配额 | 显示组织队列资源可预留配额。<br>组织管理员创建队列时预留容量受该配额限制。                                       |

- 单击“可用配额”、“组织队列资源可预留配额”列旁的。
- 修改配额。
- 单击，完成修改。

#### 步骤4 资源定价。

- 依次单击“运营 > 资源定价”。

2. 选择需要修改单价的资源页签。
3. 在需要修改定价资源的所在行，单击操作列的“修改”。  
单价列变为可编辑模式。
4. 修改单价。

#### 说明

单价需大于等于0，且最多保留两位小数。

5. 单击操作列的 ，完成定价修改。

### 步骤5 余额充值。

1. 返回组织管理界面。
2. 在目标组织所在行，单击“充值”。  
弹出充值界面，在充值界面可以查看组织的当前余额。
3. 输入充值金额和备注。

#### 说明

- 充值金额可以为负数，表示对组织余额进行扣减。
- 若组织余额小于零，组织下用户无法进行作业。

4. 单击“确定”。

**步骤6** 单击界面右上角的 ，登出当前账户。使用“组织管理员”角色的用户登录AI Space界面。

### 步骤7 创建部门。

1. 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。
2. 选择“部门”页签。
3. 单击“创建部门”。
4. 设置部门信息参数。

**表 2-5 部门信息参数说明**

| 参数   | 说明  |
|------|---|
| 基本信息 |   |
| 部门名称 | 部门的名称。<br>长度为4~32个字符，只能包含中文、字母、数字、“_”、“-”和空格，且不能全为空格。 |
| 描述   | 部门的描述信息。  |
| 启用状态 | 是否启用该部门。  |

5. 单击“确定”。

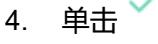
### 步骤8 部门配额。

1. 在目标部门所在行，单击“配额”。

进入配额界面，在配额界面可以查看资源配额详情。

**表 2-6 资源配额详情**

| 参数   | 说明                             |
|------|--------------------------------|
| 资源类型 | 资源的类型，包括基础资源、GPU、MIG、NPU、vNPU。 |
| 资源名称 | 资源的名称。                         |
| 已用配额 | 显示当前已使用资源数量及百分比。               |
| 可用配额 | 可使用配额数量。                       |
| 组织配额 | 显示所属组织配额。                      |

2. 单击“可用配额”列的。可用配额变为可编辑状态。
3. 修改可用配额。
4. 单击，完成修改。

**步骤9** 创建部门管理员。

1. 选择“用户”页签。
2. 单击“创建用户”。弹出创建用户对话框。
3. 输入用户信息。

**表 2-7 用户参数**

| 区域   | 参数   | 说明  |
|------|------|---|
| 基本信息 | 用户名  | 用户名只能包含大小写字母或数字。<br>可输入的字符串长度请参考安全策略的“帐号最小长度限制”和“帐号最大长度限制”。<br>默认长度范围为：6~32 |
|      | 密码   | 新建用户的密码。取值要求具体参见用户信息。   |
|      | 密码确认 | 再次输入设置的密码。  |
| 角色权限 | 角色   | 选择“DepartmentManageRole”（部门管理员）。  |
|      | 关联部门 | 从下拉列表中选择部门管理员关联部门。  |

4. 单击“确定”。

**步骤10** 创建普通用户。

1. 选择“用户”页签。

2. 单击“创建用户”。
- 弹出创建用户对话框。
3. 输入用户信息。

**表 2-8 用户参数**

| 区域   | 参数   | 说明  |
|------|------|---|
| 基本信息 | 用户名  | 用户名只能包含大小写字母或数字。<br>可输入的字符串长度请参考安全策略的“帐号最小长度限制”和“帐号最大长度限制”。<br>默认长度范围为：6~32 |
|      | 密码   | 新建用户的密码。取值要求具体参见用户信息。   |
|      | 密码确认 | 再次输入设置的密码。  |
| 角色权限 | 角色   | 选择“OrdinaryUserRole”（普通用户）。   |

4. 单击“确定”。

**步骤11** 将普通用户添加到部门。

1. 在目标部门所在行，单击“更多 > 成员管理”。
- 进入成员管理界面。
2. 单击“添加成员”，勾选需要添加的普通用户，单击“确定”。

**步骤12** 在目标成员所在行，单击操作列的“配额”，进入配额界面，在配额界面可以查看用

户的资源配置，单击  可以对用户资源配置进行修改。

----结束

# 3 操作指南

## 3.1 AI平台

### 3.2 运营

### 3.3 运维

### 3.4 系统设置

## 3.1 AI 平台

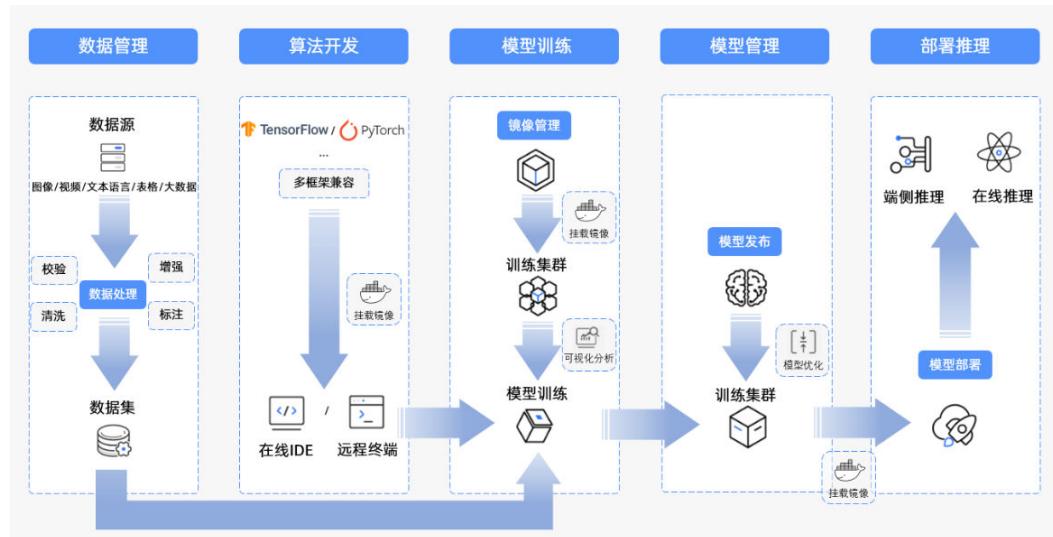
单击导航栏的“AI”，进入AI平台页面。

### 说明

- 如果用户权限不支持查看某个业务模块，则不会在左侧菜单栏中显示。
- 普通用户只能查看自己的业务信息，部门管理员可以查看该部门下所有用户的业务信息，组织管理员可以查看该组织下所有用户信息。

### 3.1.1 业务流程

图 3-1 总体流程图



## 流程说明

- **数据集管理**：用户可以上传训练数据集并进行管理。如果要将原始数据集用于进行模型训练，可在AI Space上对所上传的数据集进行如图像分类、目标检测等类型的数据标注。
- **算法开发**：为开发者提供一种在线编程的环境，该环境中包含了一些常用深度学习框架，允许开发者在线创建、编辑、调试、保存自己的算法，进而可以进行后续的模型训练工作。
- **模型训练**：使用标注完成的数据集以及开发完成的算法，进行多次反复迭代与参数调优训练，最终得到特定的结果模型。
- **模型管理**：将通过AI Space训练完成或自行上传的模型以版本形式进行存档、管理。
- **部署推理**：提供模型部署推理功能，支持在线服务和批量服务，对模型管理中的模型进行部署，支持多种深度学习框架的模型及自定义推理脚本，部署后可进行在线推理或批量推理。
- **镜像管理**：用户上传或在线制作用于训练、数据标注、可视化等操作的镜像，镜像上传至AI Space后，用户可以对镜像进行下载、分享。
- **文件管理**：为用户提供私有的文件存储空间和统一的文件管理功能，同时支持文件共享。

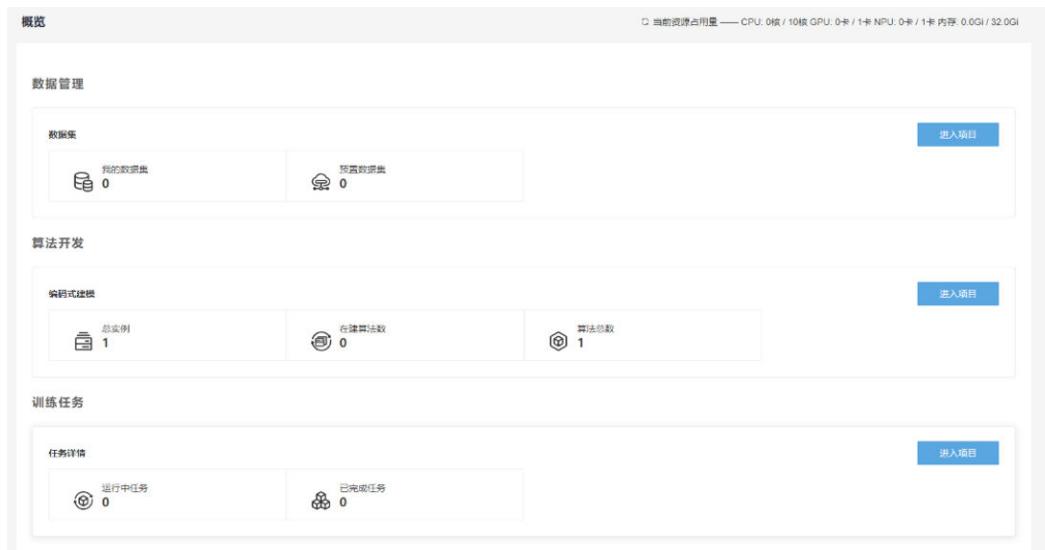
### 3.1.2 概览

“概览”界面展示了数据管理、算法开发和训练任务三个功能模块的情况，并提供快速进入各模块界面的入口。

#### 界面说明

“概览”界面各个区域的作用如表3-1所示。

图 3-2 概览界面



**表 3-1 参数说明**

| 参数   | 说明  |
|------|---|
| 数据管理 | 显示以下数据集个数： <ul style="list-style-type: none"><li>• 我的数据集</li><li>• 预置数据集</li></ul> 单击“进入项目”，跳转至数据集管理界面。                   |
| 算法开发 | 显示以下编码式建模个数： <ul style="list-style-type: none"><li>• 总实例</li><li>• 在建算法数</li><li>• 算法总数</li></ul> 单击“进入项目”，跳转至Notebook界面。 |
| 训练任务 | 显示以下训练任务的个数： <ul style="list-style-type: none"><li>• 运行中任务</li><li>• 已完成任务</li></ul> 单击“进入项目”，跳转至训练任务界面。                  |

**说明**

- 对于组织管理员，概览界面展示的内容为所属组织内所有用户创建的对应内容之和。
- 对于部门管理员，概览界面展示的内容为所属部门内所有用户创建的对应内容之和。
- 对于普通用户，概览界面展示的内容为普通用户自身创建的对应内容数量。

### 3.1.3 数据管理

#### 3.1.3.1 文件管理

##### 功能介绍

AI Space为用户提供统一的文件管理功能。文件管理界面的文件可供开展AI业务时使用，包括通过文件管理上传数据集、上传算法文件、上传镜像等。用户拥有独立的存储空间，同时AI Space支持文件共享，用户可以使用平台公用文件、组织共享文件。

文件管理界面对包括我的目录、组织目录、平台目录和用户目录，不同角色用户拥有不同权限。各目录说明如下表所示。

**表 3-2 目录说明**

| 目录   | 说明             | 后台路径  |
|------|----------------|---|
| 我的目录 | 展示当前登录用户的文件信息。 | /dfs/ai-storage/ai-prod/tenant/pvc-xxx/file-manage/user/用户名 |

| 目录   | 说明                | 后台路径  |
|------|-------------------|---|
| 组织目录 | 展示用户所属组织内共享的文件信息。 | /dfs/ai-storage/ai-prod/tenant/pvc-xxx/file-manage/org      |
| 平台目录 | 展示AI Space公用文件。   | /dfs/ai-storage/ai-prod/platform                            |
| 用户目录 | 展示组织内所有用户的目录。     | /dfs/ai-storage/ai-prod/tenant/pvc-xxx/file-manage/user/用户名 |

### □ 说明

- 用户可使用root用户登录后台，在目录对应的后台路径，对文件进行操作，操作结果将展示在文件管理界面。
- 后台路径中的pvc-xxx即PVC Volume，可参考[5.6 获取组织的PVC Volume](#)小节获取。

## 界面参数说明

文件管理界面分为左侧目录树区域和右侧文件列表区域，右侧区域参数说明如下。

表 3-3 界面参数说明

| 参数   | 说明  |
|------|---|
| 目录树  | 用户单击目录后，文件列表将展示该目录下所有文件。  |
| 文件列表 | 路径<br>展示当前目录路径。<br>用户可手动输入目录路径后按“Enter”键跳转至该目录。单击“上一级”，可跳转至上一级目录。单击路径，可跳转至指定目录。   |
|      | 名称<br>显示文件或文件夹的名称。<br>用户单击名称旁的 <span style="color: #ccc;">▼</span> ，可将文件按名称进行升序、降序排序。   |
|      | 类型<br><ul style="list-style-type: none"><li>• 文件夹</li><li>• 文件</li><li>• 图片</li><li>• 视频</li><li>• 音频</li><li>• 压缩包</li></ul> |
|      | 大小<br>显示文件的大小。  |
|      | 最近修改时间<br>显示文件最近修改的时间。<br>用户单击最近修改的时间旁的 <span style="color: #ccc;">▼</span> ，可将文件按最新修改时间进行升序、降序排序。                            |

## 文件预览

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标文件所在目录。

**步骤3** 单击需要预览的文件。

### □ 说明

- 支持预览的音频格式包括.WAV、.MP3、.FLAC，最大支持100MB。
- 支持预览的视频格式包括.MP4、.AVI、.FLV、.MOV，最大支持1GB。
- 支持预览的图片格式包括.JPG、.PNG、.TIFF、.BMP、.JPG，最大支持20MB。
- 支持预览的文本格式包括.txt、.log、.conf、.cfg、.sh、.html、.js、.py、.java、.sql、.xml、.json、.md，最大支持10MB。

**步骤4** 支持用户在线编辑文本文件。

- 用户在右上角搜索框输入关键词可对文本内容进行查找，匹配的字符会高亮展示。
- 用户在界面下方输入替换前的、替换后内容，单击“替换”，可对文本内容进行替换。

----结束

## 上传文件

对于大文件上传，上传过程中，共享存储性能可能会被上传占满，从而导致业务异常。因此，建议在上传大文件时，避免进行其他业务操作。

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标目录。

**步骤3** 单击“上传”。

**步骤4** 选择需要上传的文件。

### □ 说明

上传单个文件，文件大小最大支持1GB。

**步骤5** 单击“打开”，完成上传。

----结束

## 创建文件/文件夹

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标目录。

**步骤3** 依次单击“新建 > 新建文件/文件夹”。

弹出对话框。

**步骤4** 输入名称。

#### □ 说明

- 名称长度支持1~64位字符，支持大小写英文字母、数字、下划线（\_）、连字符（-）和英文句号（.），仅支持以数字或者字母开头。
- 文件/文件夹名称不能重复，若名称重复则将覆盖原有文件/文件夹。

**步骤5** 单击“确定”，完成创建。

----结束

## 复制文件/复制文件夹

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标文件所在目录。

**步骤3** 勾选需要复制的内容。

**步骤4** 单击界面上方的“复制到”或单击操作列的“复制到”。

弹出对话框。

**步骤5** 选择需要复制到的文件夹。

**步骤6** 单击“确定”，完成复制。

----结束

## 重命名文件/文件夹

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标文件所在目录。

**步骤3** 勾选需要重命名的内容。

**步骤4** 单击界面上方的“重命名”或单击操作列的“重命名”。

弹出对话框。

**步骤5** 输入需要修改的名称。

**步骤6** 单击“确定”。

----结束

## 剪切文件/文件夹

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标文件所在目录。

**步骤3** 勾选需要剪切的文件/文件夹。

**步骤4** 单击“更多 > 剪切到”。

弹出对话框。

**步骤5** 选择需要剪切到的目录。

**步骤6** 单击“确定”，完成剪切。

----结束

## 压缩文件

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标文件所在目录。

**步骤3** 勾选需要压缩的内容。

**步骤4** 单击界面上方的“更多 > 压缩”或单击操作列的“更多 > 压缩”。

弹出对话框。

**步骤5** 输入压缩文件参数。

表 3-4

| 参数 | 说明  |
|----|---|
| 名称 | 输入压缩文件名称。<br>支持1~60位字符，支持大小写英文字母、数字、下划线（_）、连字符（-）和英文句号（.），仅支持以数字或者字母开头。 |
| 格式 | 选择压缩文件格式，支持.tar、.tar.gz、.tar.bz2、.zip。                                  |

**步骤6** 单击“确定”，完成压缩。

----结束

## 解压压缩包

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标文件所在目录。

**步骤3** 勾选需要解压的压缩包。

**步骤4** 单击“更多 > 解压”。

### 📖 说明

- 不支持批量解压，若需要解压多个压缩包，请逐个解压。
- 若解压后的文件与解压目录已有文件同名，将覆盖原文件。

----结束

## 删除文件/文件夹

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标文件所在目录。

**步骤3** 勾选需要删除的文件/文件夹。

**步骤4** 单击“更多 > 删除”。

**步骤5** 单击“确定”，完成删除。

----结束

## 下载文件

**步骤1** 依次单击“数据管理 > 文件管理”。

进入文件管理界面。

**步骤2** 选择目标文件所在目录。

**步骤3** 勾选需要下载的文件。

**步骤4** 单击“更多 > 下载”，完成下载。

最大支持下载1GB文件。

----结束

### 3.1.3.2 数据集管理

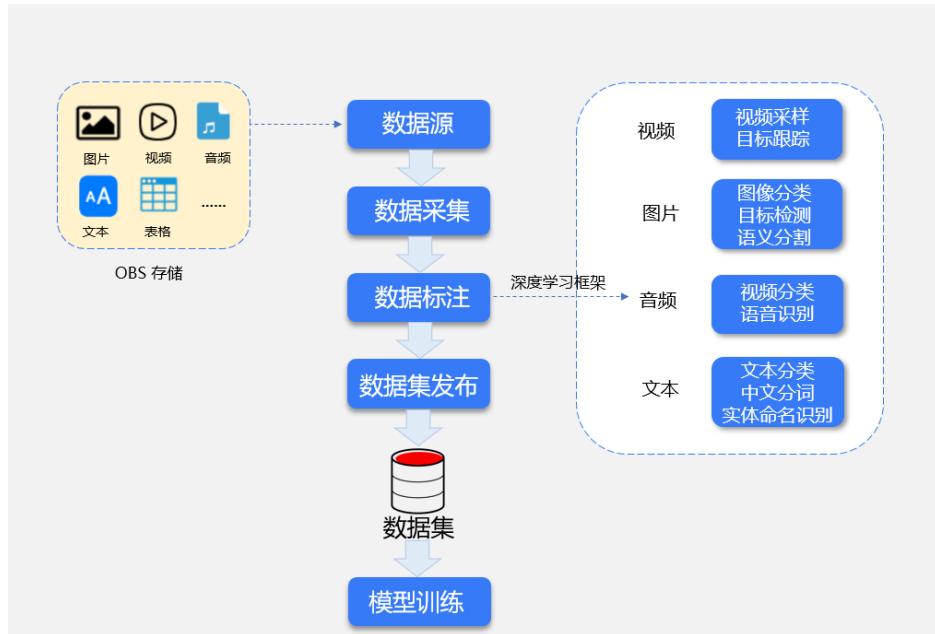
机器学习开发过程中往往需要海量数据，在通常情况下，训练数据集对于文件的质量和规格有着很高的要求。数据的质量一定程度决定了模型的好坏。

AI Space数据管理模块集成了数据导入、数据筛选、数据标注、数据增强、版本管理、标签组管理和标注服务管理等一站式数据服务，提供自动标注、数据增强等一系列数据加工方案，拥有高质量的数据标注处理算法，输出高品质的数据，从而让下游AI数据训练获得更优的训练效果。

AI Space目前支持图片、视频、文本、音频、自定义等数据类型，支持图像分类、目标检测、语义分割、目标跟踪、文本分类、中文分词、命名实体识别等数据标注功能。内置数据增强功能（针对图片类型），可以针对已有数据集进行快速扩充，获得更好的数据多样性。

## 数据集业务架构

图 3-3 数据管理业务架构



## 参数说明

表 3-5 参数说明

| 参数   | 说明  |
|------|---|
| ID   | 显示数据集的ID，ID自动递增。<br>用户可单击ID旁的 <span style="color: #0070C0;">↓</span> ，对ID进行升序、降序排序。                                     |
| 名称   | 显示数据集的名称。<br>用户可单击名称旁的 <span style="color: #0070C0;">编辑</span> 修改后单击“确定”，对数据集名称进行修改。                                    |
| 数据类型 | 显示数据集的类型。<br>用户可单击数据类型旁的 <span style="color: #0070C0;">筛选</span> ，对数据类型进行筛选。  |
| 进度   | 显示数据集标注的进度。   |
| 标注类型 | 显示数据集标注的类型，各类型数据支持的标注类型如 <a href="#">数据集标注类型</a> 所示。<br>用户可单击标注类型旁的 <span style="color: #0070C0;">筛选</span> ，对标注类型进行筛选。 |
| 状态   | 显示数据集的状态，状态说明如 <a href="#">数据集状态说明</a> 所示。<br>用户可单击状态旁的 <span style="color: #0070C0;">筛选</span> ，对状态进行筛选。               |

| 参数    | 说明  |
|-------|---|
| 当前版本  | 显示数据集的当前版本。   |
| 更新时间  | 显示数据集最新一次更新的时间。<br>用户可单击更新时间旁的  ，对更新时间进行升序、降序排序。 |
| 创建时间  | 显示数据集创建的时间。<br>用户可单击更新时间旁的  ，对创建时间进行升序、降序排序。     |
| 数据集描述 | 显示数据集的描述。   |

## 数据集标注类型

AI Space数据集目前支持视觉、语音和文本业务场景，各类型数据支持的标注类型如下表所示。

表 3-6 视觉、语音、文本业务场景

| 数据类型 | 标注类型   | 使用说明  | 智能标注 |
|------|--------|---|------|
| 图片   | 图像分类   | 对图片按标签进行分类。                                   | 部分支持 |
|      | 目标检测   | 检测图片中多个目标。                                    | 部分支持 |
|      | 语义分割   | 对图像中每一个像素点进行分类，确定每个点的类别（如属于背景、人或车等），从而进行区域划分。 | 部分支持 |
| 视频   | 目标跟踪   | 对视频采样后进行跟踪视频序列中的目标位置、信息。                      | 暂不支持 |
| 文本   | 文本分类   | 对文本按标签进行分类，表格最后还是转化为文本。                       | 部分支持 |
|      | 中文分词   | 将连续的字序列按照一定的规范重新组合成词序列的过程，表格最后还是转化为文本。        | 暂不支持 |
|      | 命名实体识别 | 识别文本中具有特定意义的实体，表格最后还是转化为文本。                   | 暂不支持 |
| 音频   | 音频分类   | 对音频按标签进行分类。                                   | 暂不支持 |
|      | 语音识别   | 将人类语音中的词汇内容转换为计算机可读的输入。                       | 暂不支持 |
| 大模型  | -      | 用于大模型微调、大模型预训练的数据集。                           | -    |
| 自定义  | -      | 导入自定义数据集。                                     | -    |

## 数据集状态说明

- 未标注：数据集所有文件均未标注。
- 手动标注中：数据集中部分文件处在“未完成”的状态，且当前未在执行“自动标注”任务。
- 自动标注中：当前数据集正在执行自动标注任务。
- 自动标注完成：数据集已完成“自动标注”任务，或“自动标注完成”后有部分文件未经“人工确认”。
- 标注完成：当前数据集所有文件都已通过“人工确认”环节。
- 目标跟踪中：对视频采样后生成的图片完成标注任务后，正在进行目标跟踪。
- 目标跟踪完成：目标跟踪完成，已生成目标ID、目标位置。
- 目标跟踪失败：文件缺失或其他原因导致失败。
- 未采样：“目标跟踪”场景下视频数据集创建完毕的初始状态。
- 采样中：“目标跟踪”场景下视频开始逐帧采样。
- 采样失败：视频采样失败。
- 数据增强中：针对图片数据集进行文件扩充，基于原始图片转换生成新图片。
- 导入中：导入文件中。
- 创建完成：大模型和自定义数据集对应状态。

## 智能标注介绍

智能标注即自动标注，通过该功能可以大幅度减少标注成本。AI Space支持以下类型的自动标注。

- 图片类：AI Space支持图像分类、目标检测和语义分割图片类自动标注。通过智能标注，算法会自动判断图像标签的置信度，并经过由用户手动确认，从而保证数据集的总体标注质量。
- 文本类：文本智能标注针对单标签类型，实现根据文本内容自动分类。

### 3.1.3.2.1 准备数据

#### ?1. 视觉类数据

视觉类数据集目前支持图片、视频两种文件类型，不同类型的标注任务对于图片和视频内容有不同的要求。

- 图像分类：识别一张图片中是否为某类物体、场景，适用于图片内容单主体，需要给整张图片分类的场景。
- 目标检测：检测图像上每个物体的位置、标签类别。适用于图片上有多个物体需要检测。
- 语义分割：语义分割是对图像中的每一个像素进行分类，目前广泛应用于医学图像与无人驾驶等。
- 目标跟踪：给定一个或多个目标，跟踪目标的移动位置变化。适用于视频文件对目标持续跟踪监测。

准备数据分为“未标注数据”和“本地已标注数据集”两类。

## 未标注数据

### 步骤1 准备标签。

在上传之前确定想要识别哪几种物体，并上传含有这些物体的图片。每个标签对应想要在图片中检测出的一种物体。

### 步骤2 准备数据。

- 保证图片质量：不能有损坏的图片；目前支持的格式包括.jpg/.png/.bmp/.jpeg；单个文件不大于5MB。
- 保证视频质量：不能有损坏的视频；目前支持的格式包括.mp4/.avi/.mkv/.mov/.webm/.wmv；单个文件不大于1024MB。
- 不要把明显不同的多个任务数据放在同一个数据集内。
- 为了保证模型的预测准确度，训练样本跟真实使用场景尽量相似。
- 为保证模型的泛化能力，数据集尽量覆盖可能出现的各种场景。

----结束

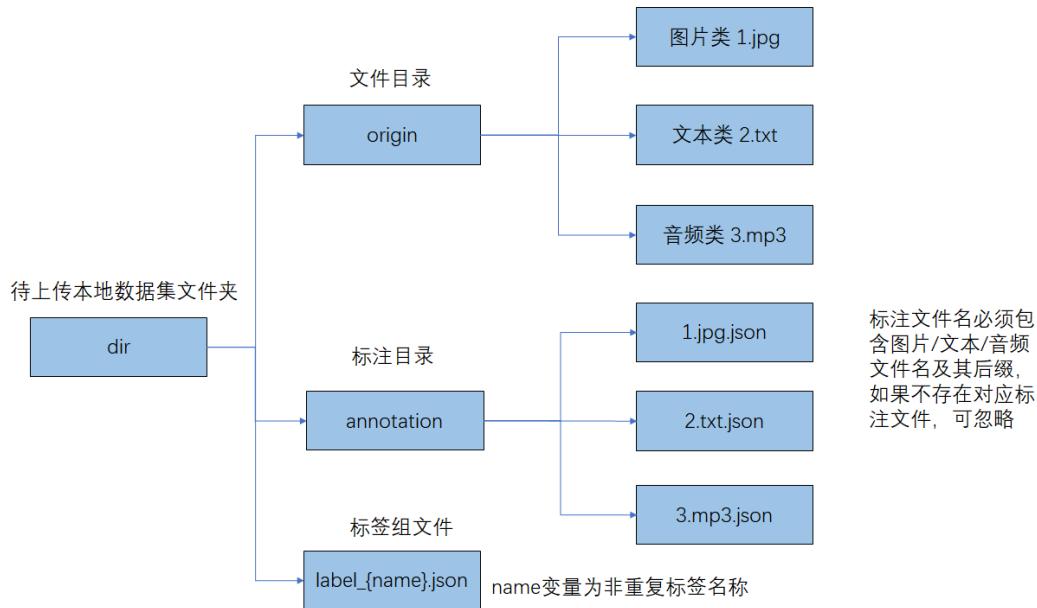
## 本地已标注数据集

- 图片格式支持.jpg/.png/.bmp/.jpeg，不大于5MB，位于origin目录下，不支持目录嵌套。
- 本地数据集需要包括源文件（origin目录）、标注文件（annotation目录）和标签文件三部分。
- 标注文件为.json格式，位于annotation目录下，标注文件名必须包含图片文件名及其后缀（如果不存在标注，可不上传），不支持目录嵌套。
- 标签文件为.json格式，命名要求为label\_{name}.json，其中name为标签组名称，不能与系统已有标签组重名。
- 导入的图片名称不能重复。

## 目录说明

本地数据集需要包括源文件（origin目录）、标注文件（annotation目录）和标签文件三部分。

图 3-4 导入数据集目录说明



## 标签格式

```
name: 名称  
color: 颜色(16进制编码)
```

详细示例：

```
[  
  {  
    "name": "行人",  
    "color": "#ffbb96"  
  },  
  {  
    "name": "自行车",  
    "color": "#fcffe6"  
  }]
```

## 标注文件

1. 图片分类：

```
name: 对应标签名称  
score: 置信分数 (0-1)
```

详细示例：

```
[{"name": "wheaten_terrier", "score": 1}]
```

2. 目标检测：

```
name: 对应标签名称  
bbox: 标注位置  
score: 置信分数 (0-1)
```

### 说明

在目标检测中，通常使用bbox ( bounding box，边界框 ) 来描述对象的空间位置。坐标以图片的左上角为原点，向右的方向为x轴正方向，向下的方向为y轴正方向，值为像素坐标。坐标的四个值分别为矩形左上角的x坐标、左上角的y坐标、右下角的x坐标和右下角的y坐标。

详细示例：

```
[{
    "name": "行人",
    "bbox": [321.6755762696266, 171.32076993584633, 185.67924201488495, 145.02639323472977],
    "score": 0.6922634840011597
},
{
    "name": "自行车",
    "bbox": [40.88740050792694, 22.707078605890274, 451.21362805366516, 326.0102793574333],
    "score": 0.6069411635398865
}]
```

## ? .2. 文本类数据

### 未标注数据

**步骤1** 准备标签。

**步骤2** 准备文本数据。

- 目前支持的格式.txt；单个文件不大于100KB，单次上传限制5000个文本文件。
- 数据集名称只支持中文、英文、数字、下划线和英文横杠。
- 如果文本数据集关联的不是预置标签组，“自动标注”功能可能无法使用。

----结束

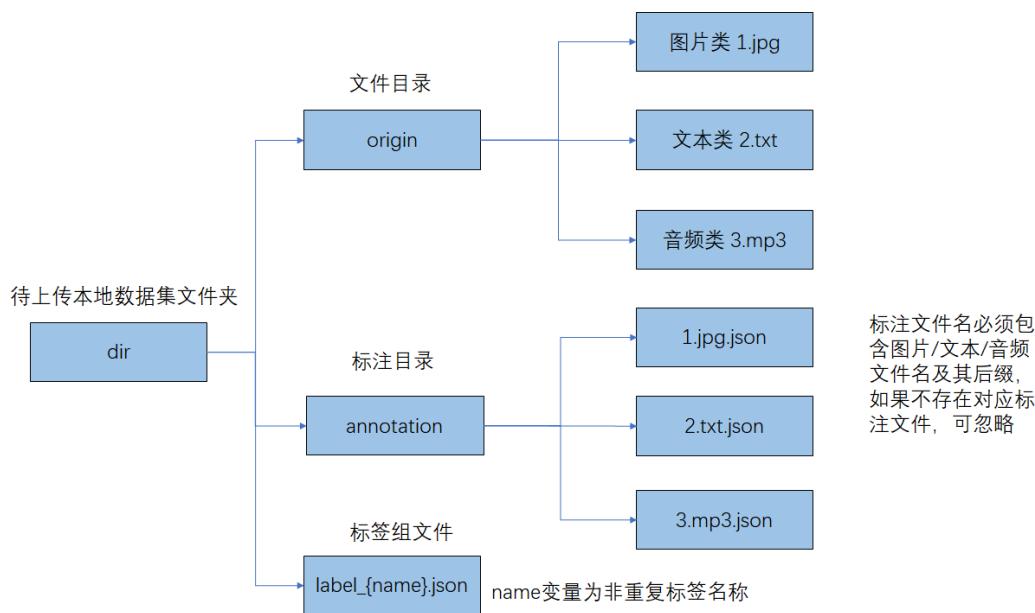
### 本地已标注数据集

- 文本格式支持.txt，位于origin目录下，不支持目录嵌套。
- 本地数据集需要包括源文件（origin目录）、标注文件（annotation目录）和标签文件三部分。
- 标注文件为.json格式，位于annotation目录下，标注文件名必须包含文本文件名及其后缀（如果不存在标注，可不上传），不支持目录嵌套。
- 标签文件为.json格式，命名要求为label\_{name}.json，其中name为标签组名称，不能与系统已有标签组重名。
- 导入的文本名称不能重复。

### 目录说明

本地数据集需要包括源文件（origin目录）、标注文件（annotation目录）和标签文件三部分。

图 3-5 导入数据集目录说明



## 标签格式

```
name: 名称
color: 颜色(16进制编码)
```

详细示例：

```
[{
  "name": "negative movie review",
  "color": "#ffbb96"
},
{
  "name": "positive movie review",
  "color": "#fcffe6"
}]
```

## 标注文件

格式：

```
name: 对应标签名称
score: 置信分数 (0-1)
```

详细示例：

```
[{"name": "negative movie review", "score": 1}]
```

## ?3. 音频类数据

### 标注类型说明

| 数据类型 | 标注类型 | 说明                  |
|------|------|---------------------|
| 音频   | 音频分类 | 音频分类是将一段音频添加到指定的标签。 |
| 音频   | 语音识别 | 语音识别是将一段音频识别成文字。    |

## 说明书

音频类型数据集暂不支持智能标注。

## 准备数据

- 数据集名称只支持中文、英文、数字、下划线和英文横杠。
- 文件格式：.mp3/.wav/.wma/.aac，单个文件不大于5MB。

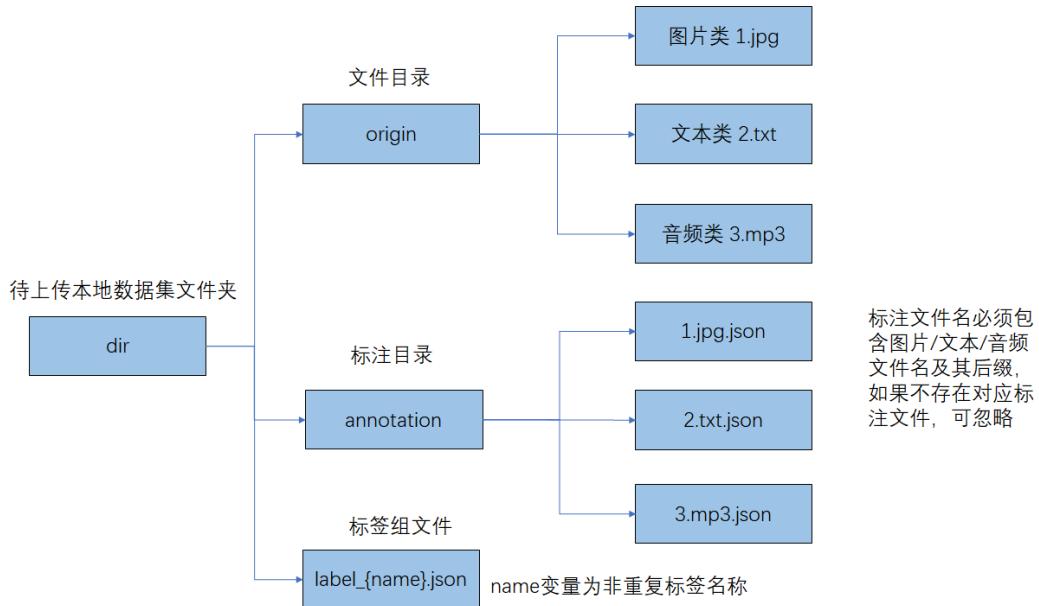
## 本地已标注数据集

- 文本格式支持.mp3/.wav/.wma/.aac，位于origin目录下，不支持目录嵌套。
- 本地数据集需要包括源文件（origin目录）、标注文件（annotation目录）和标签文件三部分。
- 标注文件为.json格式，位于annotation目录下，标注文件名必须包含文本文件名及其后缀（如果不存在标注，可不上传），不支持目录嵌套。
- 标签文件为.json格式，命名要求为label\_{name}.json，其中name为标签组名称，不能与系统已有标签组重名。
- 导入的文本名称不能重复。

## 目录说明

本地数据集需要包括源文件（origin目录）、标注文件（annotation目录）和标签文件三部分。

图 3-6 导入数据集目录说明



## 标签格式

name: 名称  
color: 颜色(16进制编码)

详细示例：

```
[{"name": "negative movie review",
```

```
        "color": "#ffbb96"
    },
{
    "name": "positive movie review",
    "color": "#fcffe6"
}]
```

## 标注文件

格式：

name: 对应标签名称  
score: 置信分数（0-1）

详细示例：

```
[{"name": "negative movie review", "score": 1}]
```

### 3.1.3.2.2 创建数据集

数据集管理模块支持以下创建数据集方法：

- 新建数据集后导入文件。
- 通过本地共享文件夹导入已有数据集。
- 通过文件管理导入已有数据集。

#### 新建数据集

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”页面。

**步骤2** 单击“创建”。

弹出“创建数据集”界面。

图 3-7 新建数据集

创建数据集

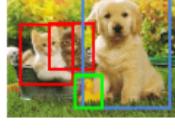
\* 创建方式  新建数据集  导入已有数据集

\* 数据集名称

\* 数据类型  图片  视频  文本  音频  大模型  自定义

\* 标注类型

图像分类 

目标检测 

语义分割 

\* 模板  单标签  多标签

\* 标签组

应用场景

数据集描述  0/100

取消  确定

**步骤3** 创建方式选择“新建数据集”。

**步骤4** 填写如下参数。

表 3-7 参数说明

| 参数    | 说明  |
|-------|---|
| 数据集名称 | 用户自定义数据集名称，数据集名称不能超过50字，支持中文、英文、数字、下划线和中划线。 |

| 参数    | 说明  |
|-------|---|
| 数据类型  | 选择数据类型，包括： <ul style="list-style-type: none"><li>图片</li><li>视频</li><li>文本</li><li>音频</li><li>大模型</li><li>自定义</li></ul>          |
| 标注类型  | 选择数据集的标注类型。   |
| 模板    | (标注类型为“图像分类”和“文本分类”时选择) 选择数据集模板，包括： <ul style="list-style-type: none"><li>单标签：单个文件只能标注单个标签。</li><li>多标签：单个文件可以标注多个标签。</li></ul> |
| 标签组   | 从下拉列表中选择标签组，中文分词、语音识别、大模型、自定义数据集不需要选择标签组。   |
| 应用场景  | 从下拉列表中选择数据集应用场景。  |
| 数据集描述 | 用户自定义数据集的描述，数据集长度不能超过100字。  |

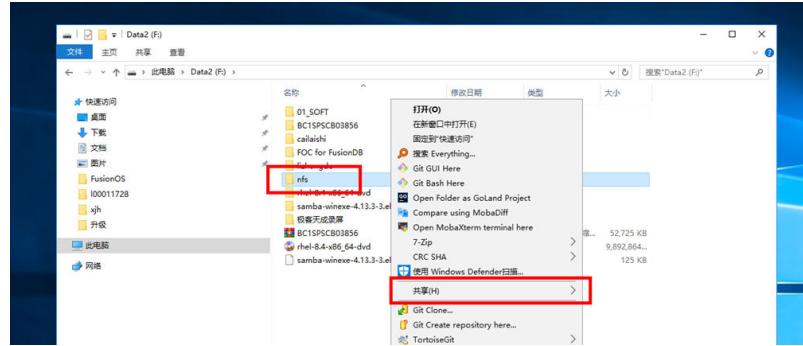
**步骤5** 单击“确定”，创建数据集成功。

----结束

## 通过本地共享文件夹导入数据集

**步骤1** 远程设置共享数据集文件夹。

图 3-8 设置共享数据集文件夹



**步骤2** 选择“数据管理 > 数据集管理”。

进入“数据集管理”页面。

**步骤3** 单击“创建”。

弹出“创建数据集”界面。

图 3-9 导入数据集

\* 创建方式  新建数据集  导入已有数据集

\* 数据集名称

\* 数据标注状态  有标注信息  无标注信息

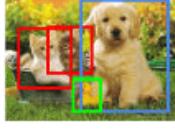
\* 数据类型  图片  文本  音频  自定义

\* 标注类型

图像分类



目标检测



语义分割



\* 数据集位置  文件管理  本地共享文件夹

\* 共享文件夹路径

\* 数据集文件夹路径

\* 用户名

\* 密码

\* IP

\* 模板  单标签  多标签

\* 标签组 ▼

标签组需要在[创建标签组](#)页面创建

数据集描述 0/100 //

**步骤4** 创建方式选择“导入已有数据集”。

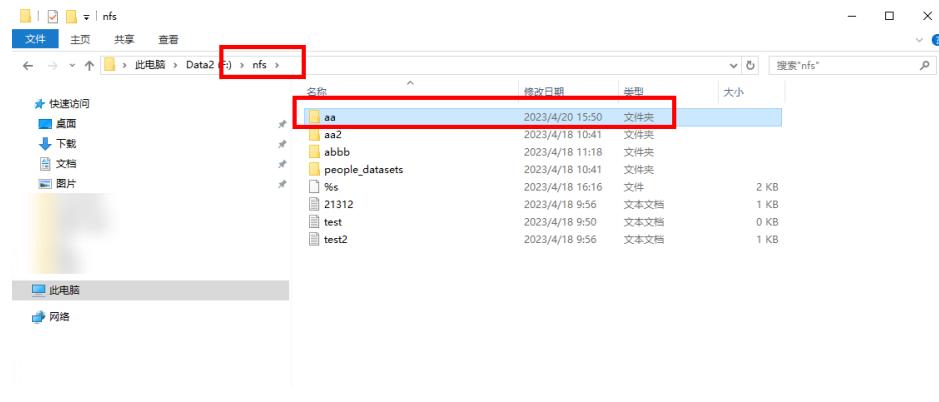
**步骤5** 填写如下参数。

表 3-8 参数说明

| 参数      | 说明  |
|---------|---|
| 数据集名称   | 用户自定义数据集名称，数据集名称不能超过50字。  |
| 数据标注状态  | 选择数据标注状态，包括：<br>● 有标注信息<br>● 无标注信息  |
| 数据类型    | 选择数据类型，包括：<br>● 图片<br>● 文本<br>● 音频<br>● 自定义                                       |
| 标注类型    | 选择数据集的标注类型。   |
| 数据集位置   | 选择“本地共享文件夹”。  |
| 共享文件夹路径 | 填写共享文件夹的路径。   |
| 数据文件夹路径 | 填写数据文件夹的路径。   |
| 用户名     | 填写用户名。  |
| 密码      | 填写密码。   |
| IP      | 填写IP地址。   |
| 模板      | (标注类型为“图像分类”和“文本分类”时选择) 选择数据集模板，包括：<br>● 单标签：单个文件只能标注单个标签。<br>● 多标签：单个文件可以标注多个标签。 |
| 标签组     | (数据标注状态为“无标注信息”时选择) 从下拉列表中选择标签组。中文分词、语音识别、自定义数据集不需要选择标签组。                         |
| 数据集描述   | 用户自定义数据集的描述，数据集长度不能超过100字。  |

共享文件夹路径、数据文件夹路径示例：

图 3-10 共享文件夹路径、数据文件夹路径示例



创建数据集



步骤6 单击“确定”，创建数据集成功。

图 3-11 数据集创建完成



----结束

## 通过文件管理导入数据集

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”页面。

**步骤2** 单击“创建”。

弹出“创建数据集”界面。

**步骤3** 创建方式选择“导入已有数据集”。

**步骤4** 填写如下参数。

**表 3-9** 参数说明

| 参数     | 说明   |
|--------|--|
| 数据集名称  | 用户自定义数据集名称，数据集名称不能超过50字。   |
| 数据标注状态 | 选择数据标注状态，包括： <ul style="list-style-type: none"><li>● 有标注信息</li><li>● 无标注信息</li></ul>   |
| 数据类型   | 选择数据类型，包括： <ul style="list-style-type: none"><li>● 图片</li><li>● 文本</li><li>● 音频</li><li>● 自定义</li></ul>                            |
| 标注类型   | 选择数据集的标注类型。  |
| 数据集位置  | 选择“文件管理”。  |
| 文件管理   | 单击“文件管理”，弹出文件管理界面。<br>选择数据集文件夹，单击“确定”。   |
| 模板     | (标注类型为“图像分类”和“文本分类”时选择)选择数据集模板，包括： <ul style="list-style-type: none"><li>● 单标签：单个文件只能标注单个标签。</li><li>● 多标签：单个文件可以标注多个标签。</li></ul> |
| 标签组    | (数据标注状态为“无标注信息”时选择)从下拉列表中选择标签组。中文分词、语音识别、自定义数据集不需要选择标签组。   |
| 数据集描述  | 用户自定义数据集的描述，数据集长度不能超过100字。   |

**步骤5** 单击“确定”，创建数据集成功。

----结束

## 导入文件

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 单击操作列的“导入”。

**步骤3** 选择导入文件方式，包括从本地上传文件和从文件管理上传文件。

- 选择“上传文件”：单击“上传文件”，从本地选择文件上传。
- 选择“文件管理”：单击“文件管理”，弹出文件管理弹窗，勾选需要导入的文件/文件夹，单击“确定”。

### 说明

- 文件和文件夹不能让同时选择。
- 选择当前文件夹，只针对当前文件夹下的文件进行上传，不含子文件夹内的文件。

**步骤4** 单击“确定”，完成导入文件。

----结束

## 删除文件

**步骤1** 选择“数据管理 > 数据集管理”。

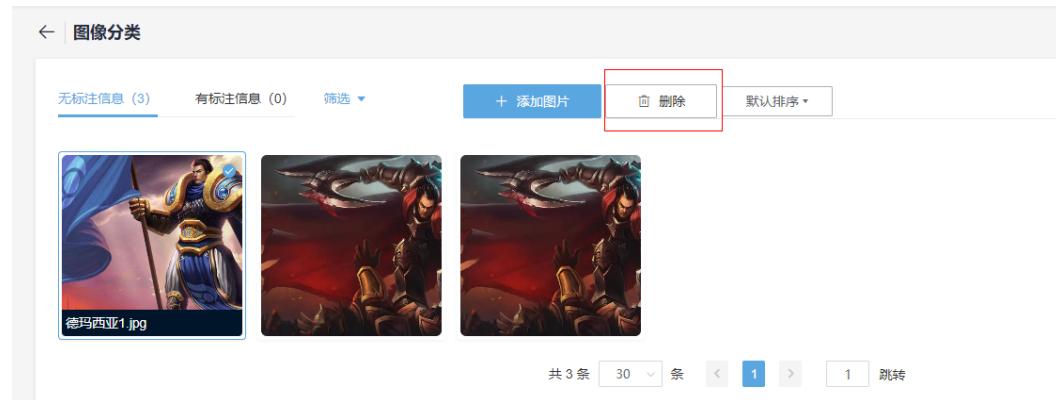
进入“数据集管理”界面。

**步骤2** 单击数据集名称或操作列的“查看与标注”。

进入数据集详情页。

**步骤3** 单击文件右上角的圆圈，勾选需要删除的数据。

图 3-12 删除文件



**步骤4** 单击“删除”。

弹出提示框。

**步骤5** 单击“确定”。

----结束

### 3.1.3.2.3 数据标注

AI Space数据管理模块提供了数据标注功能。

数据集详情页分别展示了标注“无标注信息”和“有标注信息”的文件，单击文件，即可进行文件的预览，对于已标注文件，左上角会显示该文件的标签信息和标注方式。

#### ?.1. 自动标注

自动标注结合已构建模型进行自动化标注，快速完成标注操作。数据管理模块提供了自动标注功能，为开发者节省70%以上的标注时间。AI Space支持对以下标注类型的自动标注。

表 3-10 自动标注支持标注类型

| 标注类型 | 数据类型 |
|------|------|
| 图像分类 | 图片   |
| 目标检测 | 图片   |
| 语义分割 | 图片   |
| 文本分类 | 文本   |

#### □ 说明

在自动标注前，请确保准备好标注算法和类型，详情请参见[标注算法和模型](#)。

## 操作步骤

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 在待需要自动标注的数据集所在行，单击“更多 > 自动标注”。

弹出“自动标注”提示框。

图 3-13 自动标注



**步骤3** 选择合适的标注服务。

**步骤4** 单击“高级设置”，展开高级设置栏。

**步骤5** 根据文件标注信息选择需要标注的数据“无标注信息/有标注信息”。

**步骤6** 单击“确定”。

数据集进入“自动标注中”的状态。

----结束

## 停止标注

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 在状态为“自动标注中”的数据集所在行，单击“停止”。

弹出提示框。

**步骤3** 单击“确定”。

----结束

## 重新自动标注

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 在已完成自动标注的数据集所在行，单击“自动标注”。

弹出“自动标注”提示框。

**步骤3** 根据模型类型选择标注服务和高级设置。

**步骤4** 单击“确定”，重新进行自动标注。

所有标注信息都会被清除，并再次运行自动标注算法生成新的结果。

----结束

## ?2. 手动标注

### ?1. 图像分类

#### 操作步骤

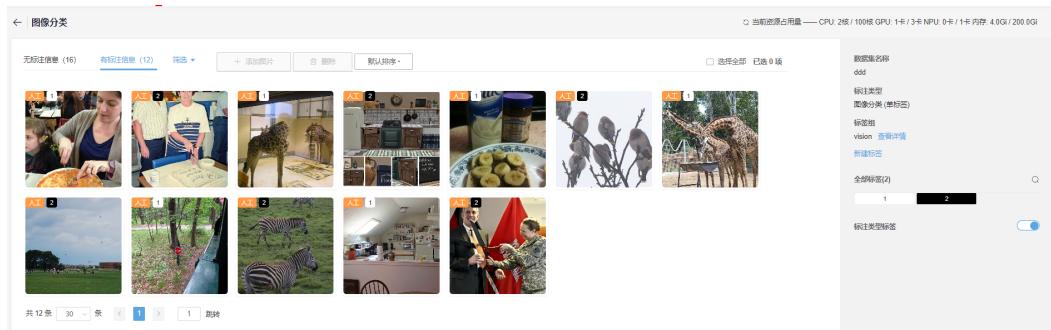
**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

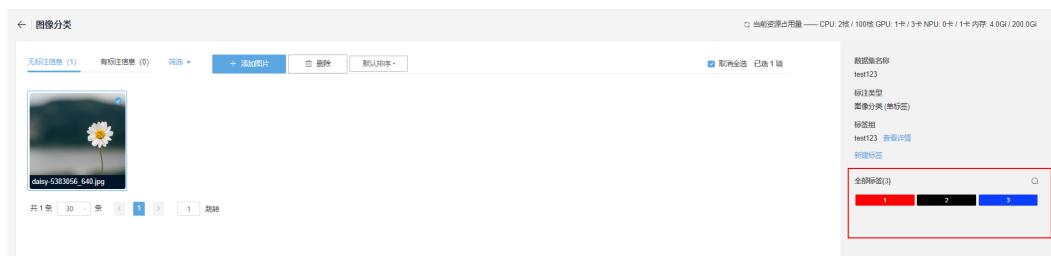
**步骤3** 单击文件右上角的圆圈，勾选需要标注的数据。

图 3-14 手动标注



**步骤4** 根据数据集内容选择需要标注的标签，单击“全部标签”中的标签名进行标注。

图 3-15 手动标注效果



**步骤5** 对于支持多标签的标注类型，创建数据集时选择多标签，标注时可以根据图片内容选择多个标签。

图 3-16 多标签标注



----结束

## ? .2. 目标检测

### 操作步骤

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

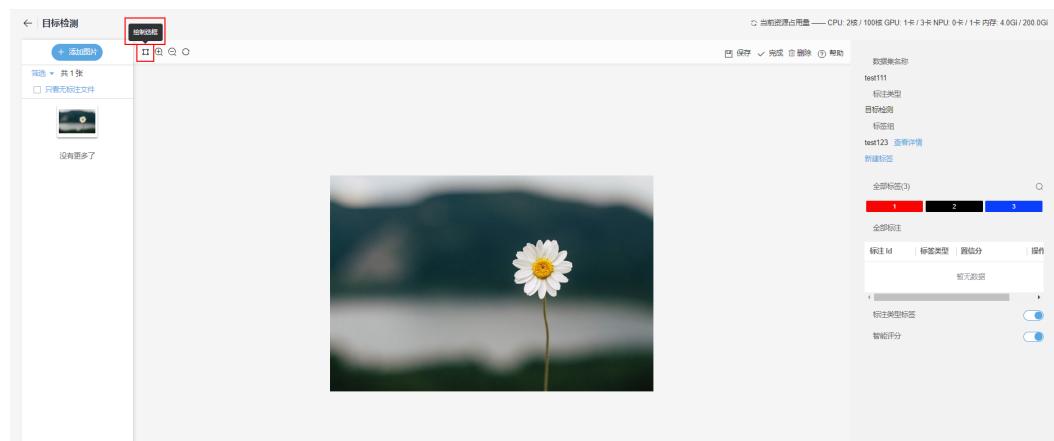
**步骤2** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

**步骤3** 单击文件右上角的圆圈，勾选需要标注的数据。

**步骤4** 单击“去标注”进入图片标注界面。

**步骤5** 单击“绘制选框”图标，用鼠标框选图片中的物体所在区域。

图 3-17 绘制标注框

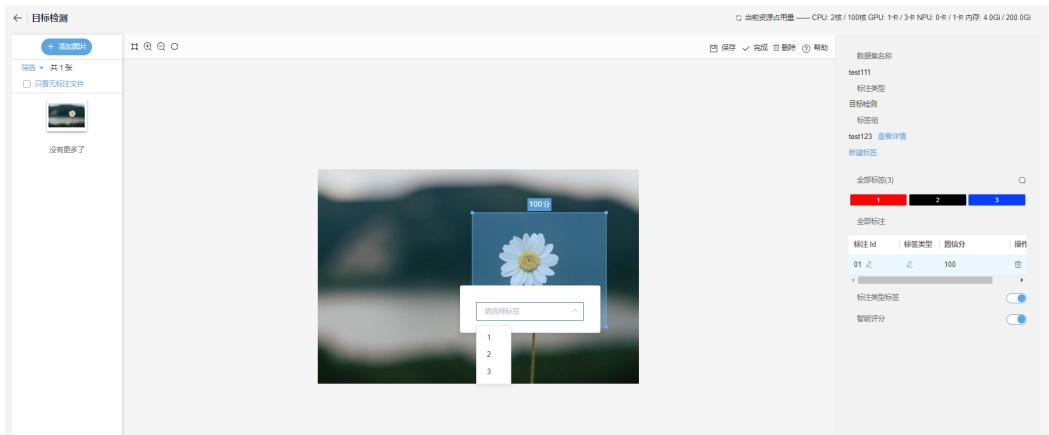


## 📖 说明

- 拖动标注框顶点可以调节大小，拖动标注框的内部可以调节位置。
- 目标检测仅支持矩形标注框。
- 按键盘的左右方向键，选择切换图片，重复上述操作继续进行图片标注。
- 在标注窗口中滚动鼠标放大或缩小图片，方便快速定位到物体位置。

**步骤6** 在弹出的对话框中选择标签。

**图 3-18 选择标签**



## 📖 说明

同一张图片可以增加多个标签，且不同的标签可以设置不同的颜色，方便识别。

**步骤7** 标注完成后单击“保存”和“完成”。

----结束

## ? .3. 语义分割

### 操作步骤

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

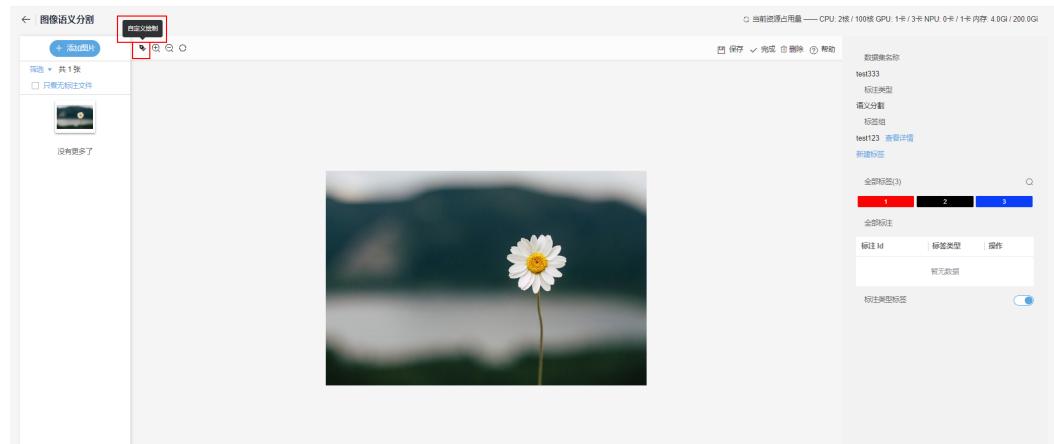
**步骤2** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

**步骤3** 单击文件右上角的圆圈，勾选需要标注的数据。

**步骤4** 单击“去标注”进入标注页面。

**步骤5** 单击画布左上角“自定义绘制”图标或快捷键N进行绘制。

图 3-19 语义分割

**说明**

- 按键盘ESC键放弃绘制，按键盘F键快速完成绘制。
- 标注绘制完成按键盘Enter键快速提交，鼠标左键拖动整体可以实现标注位移。
- 继续标注可以使用键盘Left键上一张、Right键下一张。
- 为了更加快捷的进行绘制，按住键盘SHIFT键拖动鼠标可以进行快速绘制。
- 在标注的过程中出现标注错误，按下键盘Backspace键可以进行快速删除。
- 标注完成后，对不满意的地方选中标注点拖拽进行标注点位置修改。
- 拖拽标注点在增长距离的同时可以增加标注点，单击小的标注点将标注点实体化。
- 单击实心标点可以删除标注点。

**步骤6 选择标签。**

**步骤7 标注完成后单击“保存”和“完成”。**

----结束

**?4. 目标跟踪****操作步骤**

**步骤1 选择“数据管理 > 数据集管理”。**

进入“数据集管理”界面。

**步骤2 单击操作列的“导入”，导入视频。**

数据集自动进入“采样中”，采样完成后数据集的状态变成“未标注”状态。

图 3-20 采样中

| 数据集管理                    |    |        |      |      |      |        |                     |                     |       |               |
|--------------------------|----|--------|------|------|------|--------|---------------------|---------------------|-------|---------------|
| 操作                       |    | 名称     |      | 数据类型 | 进度   | 标注类型   | 状态                  | 当前版本                | 更新时间  | 数据集描述         |
| 我的数据集                    |    |        |      |      |      |        |                     |                     |       |               |
| <input type="checkbox"/> | ID | 名称     | 数据类型 | 进度   | 标注类型 | 状态     | 当前版本                | 更新时间                | 数据集描述 | 操作            |
| <input type="checkbox"/> | 79 | test03 | 视频   | 0%   | 目标跟踪 | 采样中    | 2023-06-14 10:25:06 | 2023-06-14 10:25:06 | 待标注   | 停止 置顶 修改      |
| <input type="checkbox"/> | 78 | test01 | 图片   | 0%   | 语义分割 | 未标注    | 2023-06-14 10:19:06 | 2023-06-14 10:19:06 | 待标注   | 导入 查看与标注 更多 ▾ |
| <input type="checkbox"/> | 77 | test   | 图片   | 0%   | 目标检测 | 未标注    | 2023-06-14 10:15:59 | 2023-06-14 10:15:59 | 待标注   | 导入 查看与标注 更多 ▾ |
| <input type="checkbox"/> | 76 | ast    | 图片   | 0%   | 图像分类 | 自动标注完成 | 2023-06-13 15:56:40 | 2023-06-13 15:56:40 | 已标注   | 发布 导入 更多 ▾    |
| <input type="checkbox"/> | 70 | 11     | 图片   | 0%   | 图像分类 | 自动标注完成 | 2023-06-13 15:56:01 | 2023-06-13 15:56:01 | 已标注   | 发布 导入 更多 ▾    |
| <input type="checkbox"/> | 75 | rlead  | 图片   | 0%   | 图像分类 | 自动标注中  | 2023-06-12 15:38:21 | 2023-06-12 15:38:21 | 待标注   | 停止 置顶 修改      |

**图 3-21 采样完成**

The screenshot shows the 'My Dataset' section of the Fusion AI Space interface. It displays a table with one row of data. The columns include: ID (79), 名称 (test03), 数据类型 (Video), 进度 (0%), 标注类型 (目标跟踪), 状态 (未标注), 当前版本 (2023-06-14 10:25:06), and 操作 (Import, View & Annotation, More). A red box highlights the 'View & Annotation' button in the '操作' column.

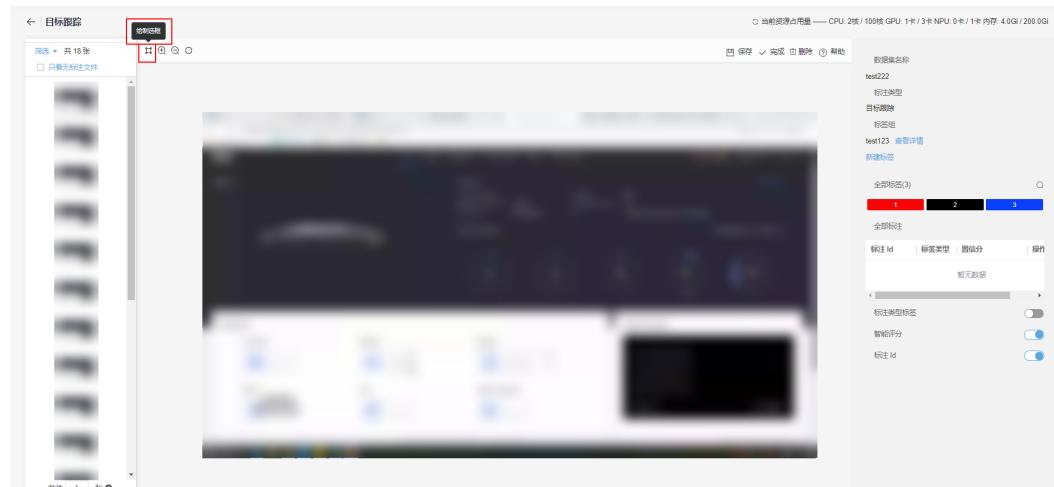
| ID | 名称     | 数据类型 | 进度 | 标注类型 | 状态  | 当前版本                | 更新时间 | 数据集描述 | 操作            |
|----|--------|------|----|------|-----|---------------------|------|-------|---------------|
| 79 | test03 | 视频   | 0% | 目标跟踪 | 未标注 | 2023-06-14 10:25:06 |      |       | 导入 查看与标注 更多 ▾ |

**步骤3** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

**步骤4** 单击文件右上角的圆圈，勾选需要标注的数据。

**步骤5** 单击“去标注”进入标注页面。

**步骤6** 单击左上角的“绘制选框”，选择目标图标进行标注。

**图 3-22 手动标注**

### 说明

单击标注框可以进行位置、大小的调节。

**步骤7** 在弹出的对话框中选择标签。

### 说明

标注完成后可以在右下角的标注ID处进行标注ID的修改。

**步骤8** 标注完成后，单击右上角“保存”，进行保存。

**步骤9** 单击“完成”，跳转到下一张继续标注。

----结束

## ? .5. 文本标注

文本标注包括文本分类标注、中文分词标注和命名实体组织标注。

## 文本分类

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

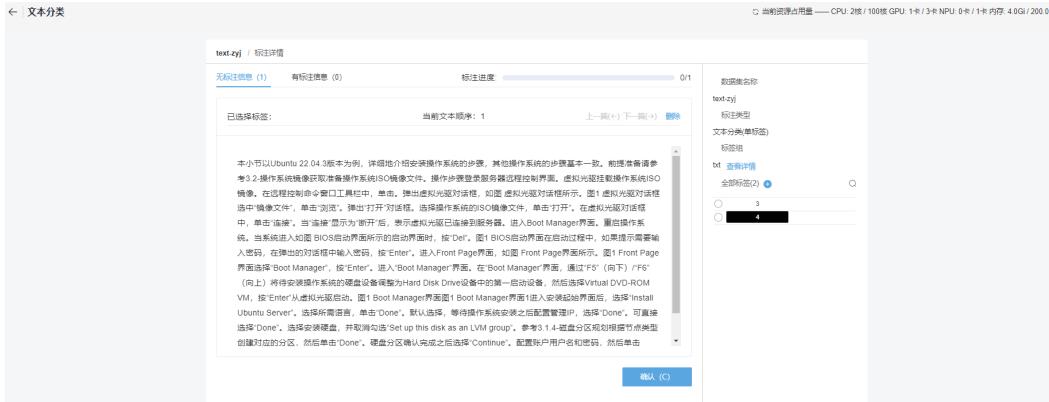
**步骤3** 勾选需要标注的数据。

图 3-23 文本列表



**步骤4** 单击“去标注”进入标注详情界面。

图 3-24 文本标注详情



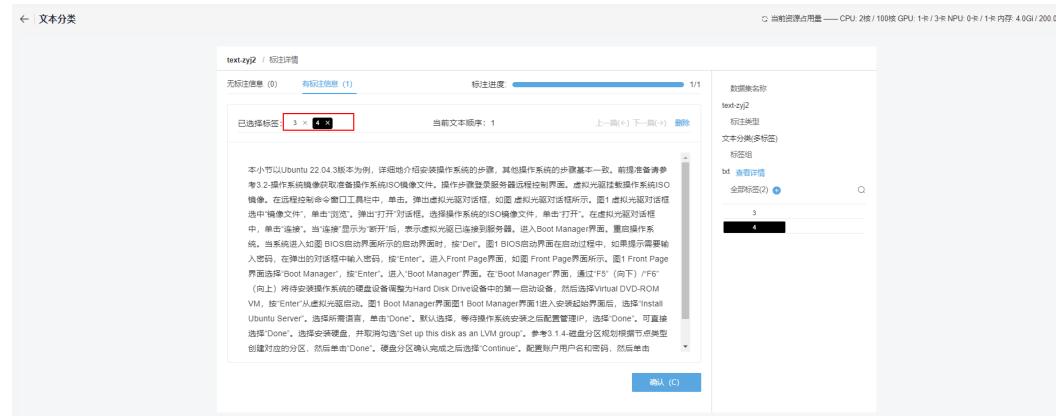
**步骤5** 根据文本内容选择标签进行标注，单击“确认”或使用键盘C键，手动标注完成。

图 3-25 手动标注完成



“文本分类”数据集支持“多标签”的标注，创建数据集时选择“多标签”，根据文本内容可以选择多个标签。

图 3-26 文本分类多标签



----结束

## 中文分词

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

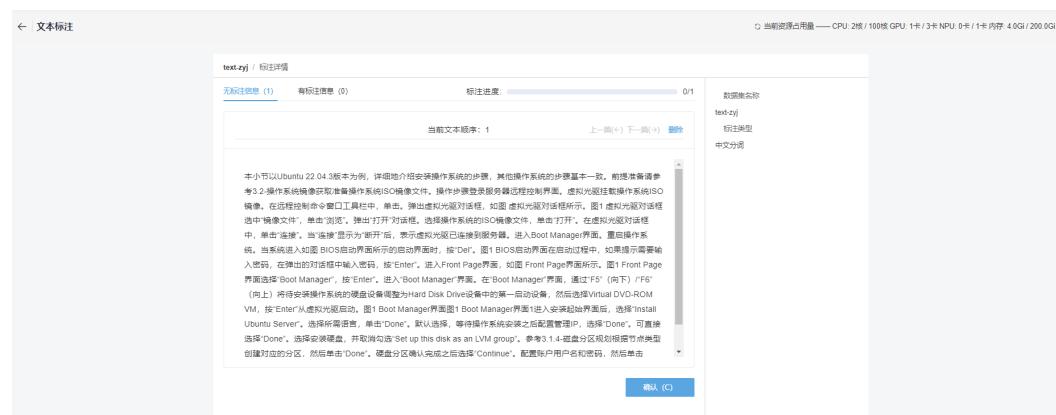
**步骤3** 勾选需要标注的数据。

图 3-27 文本列表



**步骤4** 单击“去标注”进入标注详情页，根据文本内容进行分词标注。

图 3-28 中文分词标注详情



**步骤5** “中文分词”手动标注完成数据展示。

图 3-29 手动标注完成



----结束

## 命名实体识别

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

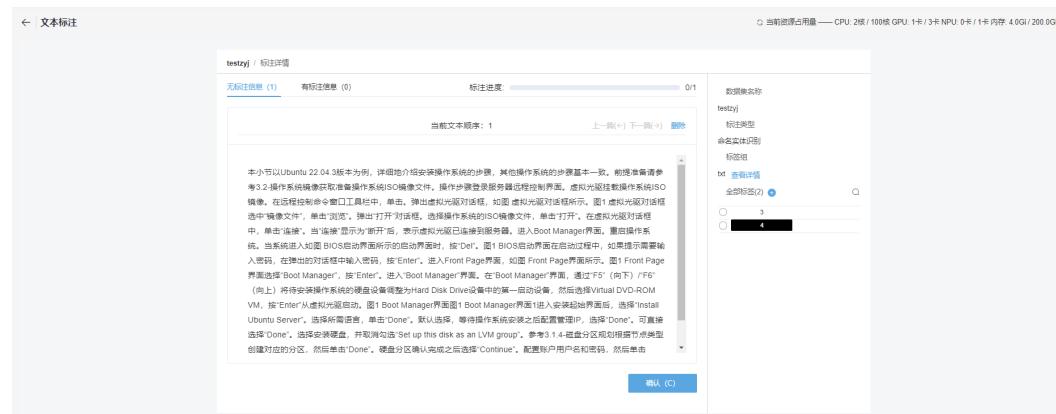
**步骤3** 勾选需要标注的数据。

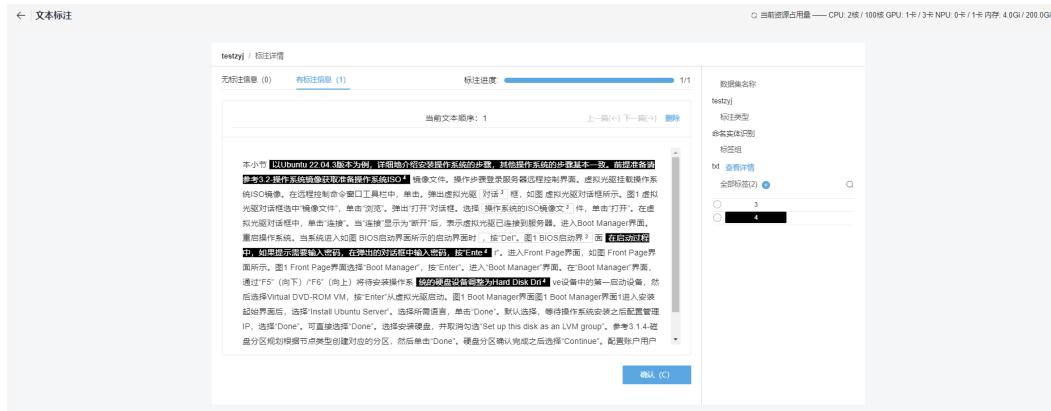
图 3-30 命名实体识别列表



**步骤4** 单击“去标注”进入标注详情页，根据文本内容进行标注。

图 3-31 文本预览



**步骤5** “命名实体识别” 手动标注完成数据展示。**图 3-32 命名实体识别****----结束****? .6. 音频标注**

音频标注包括音频分类标注和语音识别标注。

**音频分类**

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

**步骤3** 勾选需要标注的数据。

**步骤4** 单击“去标注”进入标注详情页。

**步骤5** 听取音频文件，选择标签，单击“确定”或使用快捷键C标注完成。

**----结束****语音识别**

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 单击数据集名称或操作列的“查看与标注”，进入数据集详情页。

**步骤3** 勾选需要标注的数据。

**步骤4** 单击“去标注”进入标注详情页。

**步骤5** 听取音频文件，在输入框内填入音频内容或标签，单击“确定”或快捷键C标注完成。

**----结束****3.1.3.2.4 数据增强**

一个优秀的深度学习训练模型通常需要海量数据，如果开发者的原始数据集达不到一定量级，就很难训练出具有泛化能力的模型。基于数据增强的数据集扩容，可一定程

度上缓解此类问题。AI Space提供了图像层面的增强方法，基于原始数据集单张图片进行转换操作，从而达成对数据集的扩充效果。

## 前提条件

- 数据集状态必须为“自动标注完成”或“标注完成”。
- 数据增强操作只支持“图像分类”和“目标检测”标注类型的数据集。
- 数据增强操作只针对原始图片进行转换。
- 数据增强可以针对同一批原始图片进行多种转换。
- 使用导入已有数据集方式创建的数据集暂不支持数据增强功能。

## 增强类型说明

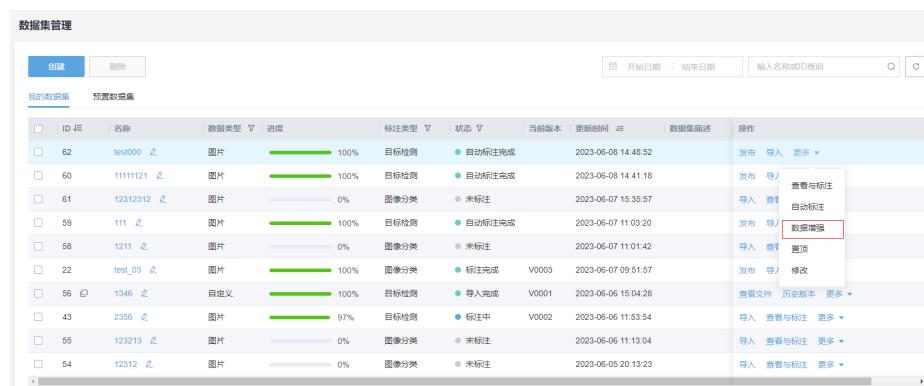
AI Space支持多种增强类型算法，包括去雾、增雾、对比度增强和直方图均衡化。

| 增强类型          | 名称     | 说明                |
|---------------|--------|-------------------|
| dehaze        | 去雾     | 对单张图像数据进行去雾操作。    |
| addhaze       | 增雾     | 对单张图像数据进行增雾操作。    |
| ACE           | 对比度增强  | 对单张图像数据进行去雾对比度增强。 |
| HIST_equalize | 直方图均衡化 | 对单张图像数据进行直方图均衡化。  |

## 操作步骤

- 步骤1 选择“数据管理 > 数据集管理”。
- 进入“数据集管理”界面。
- 步骤2 在待增强的数据集所在行，单击“更多 > 数据增强”。
- 弹出“数据增强”提示框。

图 3-33 数据增强



- 步骤3 选择需要增强的类型（支持多选），单击“确定”。

图 3-34 选择增强类型

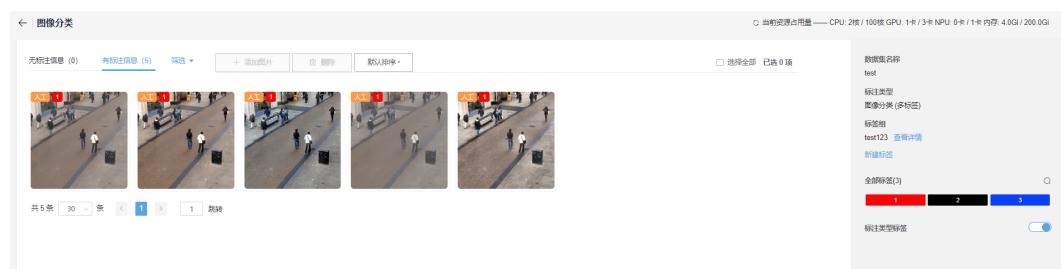


**步骤4** 当前数据集状态变更为“数据增强中”，等待增强完成即可。

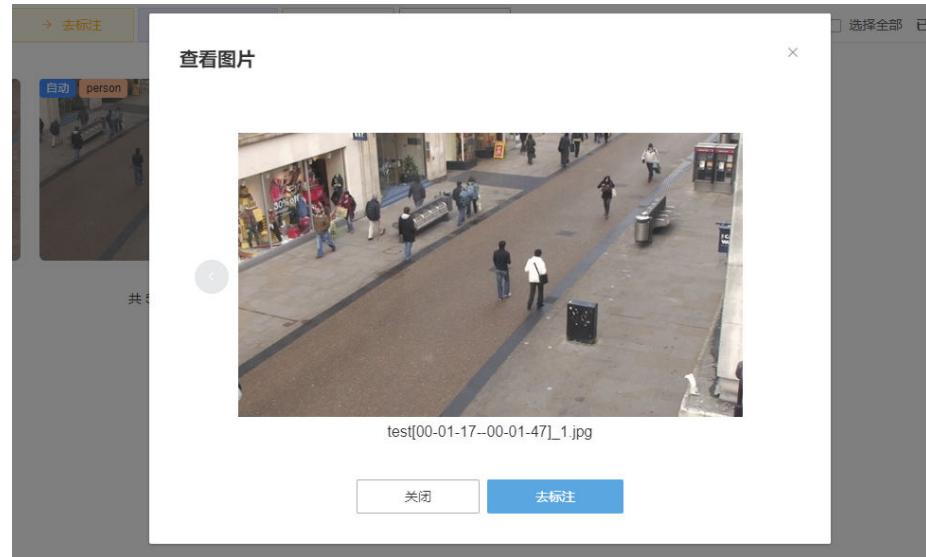
**步骤5** 查看增强后的数据集。

单击“查看与标注”，进入数据集详情页。

图 3-35 图像分类数据增强结果



**步骤6** 增强效果对比。

**图 3-36** 数据集原图展示

数据集“去雾”效果展示。

**图 3-37** 去雾

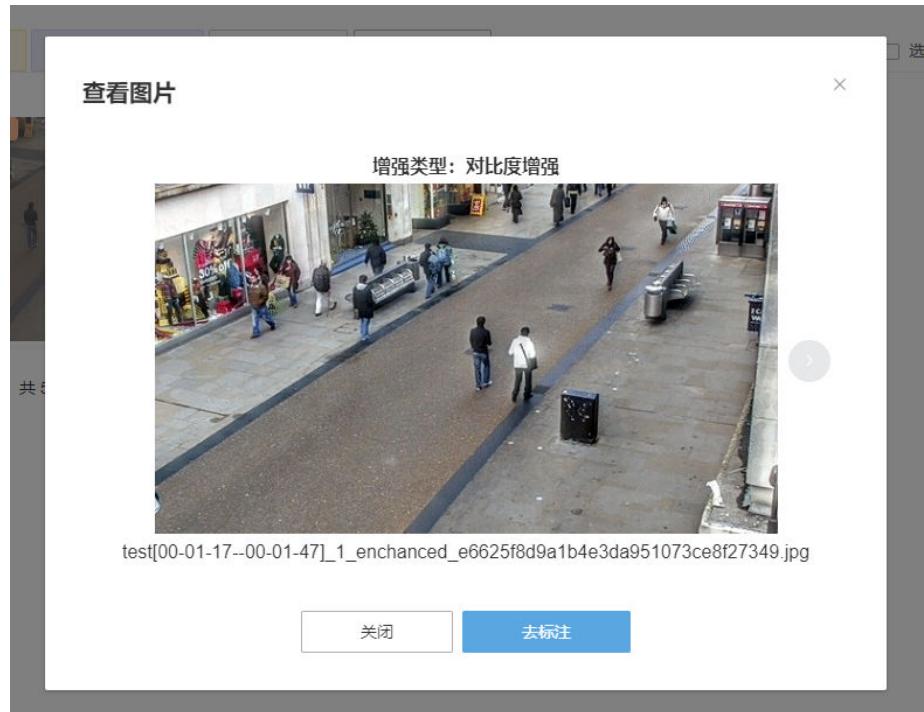
数据集“增雾”效果展示。

图 3-38 增雾



数据集“对比度增强”效果展示。

图 3-39 对比度增强



数据集“直方图均衡化”效果展示。

图 3-40 直方图均衡化



----结束

### 3.1.3.2.5 发布数据集

AI Space提供基于不同版本进行内容区分的数据集管理功能。后续模型训练阶段可选择不同版本的数据集进行训练、开发。开发者在完成数据集的标注之后，就可以对数据集进行发布操作，生成此数据集的一个新版本。

#### 前提条件

当前数据集已经标注完成（自动标注完成，或者手动标注完成均可）。

#### 关于数据集版本

- 新创建的数据集（未发布过），无数据集版本信息，只有执行发布之后，才能用于模型训练或开发。
- 数据集版本统一按照V0001, V0002, ... 规则命名。
- AI Space支持多种格式数据的加载。
- 开发者可以指定特定版本作为数据集当前版本。
- 针对每个数据集版本，可以查看当前数据集文件数量、标注进度和数据集状态。

#### 发布数据集

**步骤1** 选择“数据管理 > 数据集管理”。

进入“数据集管理”界面。

**步骤2** 在状态为“标注完成”的数据集所在行，单击“发布”。

弹出“发布数据集”提示框。

**图 3-41 版本发布**

| ID | 名称       | 数据类型 | 进度   | 标注类型 | 状态     | 当前版本                | 更新时间 | 数据集描述 | 操作          |
|----|----------|------|------|------|--------|---------------------|------|-------|-------------|
| 62 | test000  | 图片   | 100% | 目标检测 | 自动标注完成 | 2023-06-08 14:46:52 |      |       | 发布 导入 更多    |
| 60 | 11111121 | 图片   | 100% | 目标检测 | 自动标注完成 | 2023-06-08 14:41:18 |      |       | 发布 导入 更多    |
| 61 | 12312312 | 图片   | 0%   | 图像分类 | 未标注    | 2023-06-07 15:35:57 |      |       | 导入 查看与标注 更多 |

**步骤3 设置数据集信息。****说明**

导出格式仅支持视觉类型数据集。

**图 3-42 发布数据集**

发布数据集

数据集名称 1

当前版本 无

\* 下一版本 V0001

导出格式 JSON

版本描述 请输入内容 0/100

取消 确定

**步骤4 单击“确定”。****----结束****数据集版本管理****步骤1** 数据集发布后，单击“更多 > 历史版本”，进入“数据集版本管理”界面，查看此数据集的所有版本信息。

图 3-43 数据集管理

The screenshot shows the 'Data Set Management' interface. At the top, there are buttons for 'Create' and 'Delete'. Below that is a search bar with filters for 'Start Date' and 'End Date' and a search input field. The main area is titled 'My Data Sets' and contains a table of datasets. The columns include ID, Name, Data Type, Progress, Annotation Type, Status, Current Version, Update Time, Description, and Operations. A specific row for dataset ID 56 is highlighted with a red box around the 'History Version' link in the operations column. The table shows 42 items with page navigation at the bottom.

步骤2 在“数据集版本管理”界面，单击“详情”，查看版本详细信息。

图 3-44 自定义数据集版本详情

The screenshot shows the 'Custom Data Set Version Management' interface. It displays a table with columns: ID, Name, Data Type, Annotation Type, Export Format, Is Current Version, Version Number, and Creation Time. A specific row for dataset ID 12 is selected. To the right, a detailed view panel shows information for 'test(V0001)'. It includes fields for Status (标注完成), File Count (文件数量), and Annotation Progress (标注进度). There are also links for 'Details', 'View File', and 'More'.

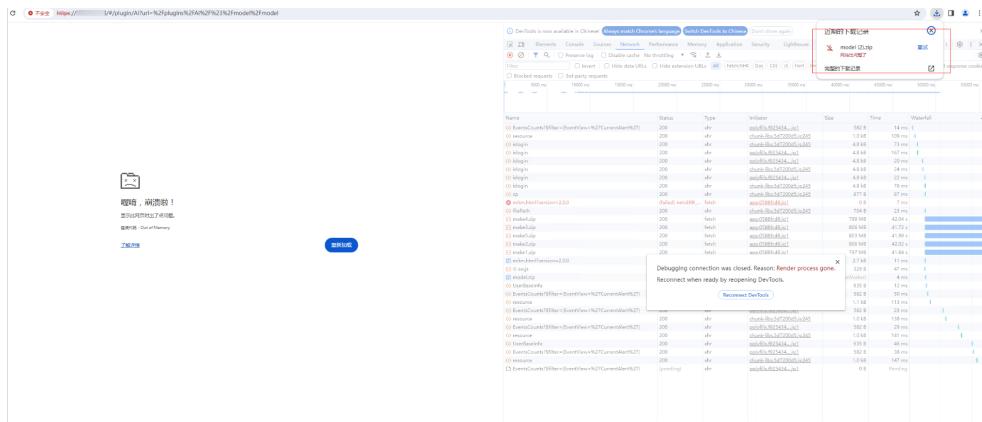
图 3-45 非自定义数据集版本详情

The screenshot shows the 'Non-Custom Data Set Version Management' interface. It displays a table with columns: ID, Name, Data Type, Annotation Type, Export Format, Is Current Version, Version Number, and Creation Time. A specific row for dataset ID 9 is selected. To the right, a detailed view panel shows information for '12314(V0001)'. It includes fields for Status (标注完成), File Count (文件数量), and Annotation Progress (标注进度). There are also links for 'Details', 'View Annotation', 'Generate Predefined Dataset', and 'More'.

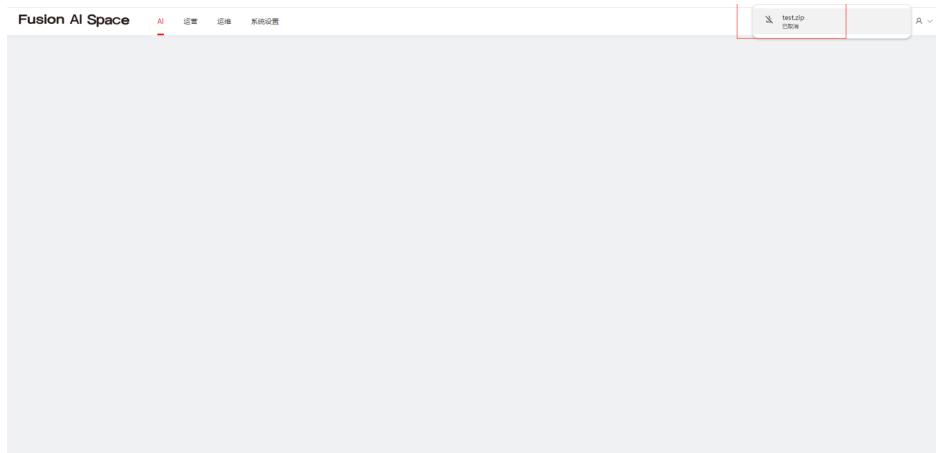
步骤3 单击“更多 > 导出”，可以导出数据集至本地。

## 说明

- 导出数据集并解压后，若需要查看annotation文件夹中的JPG格式文件，请以文本方式打开。
- 若数据集名称中包含中文字符，则导出数据集至本地可能存在报错。
- 在下载过程中，如果如下图所示，出现界面崩溃情况，这可能与用户的个人设备相关，可以由超级管理员登录后台进行下载。若需要进一步的帮助，请参照[7 如何获取帮助](#)章节，联系技术支持团队。



- 下载文件时，如下图所示，可能会遭遇会话超时的问题，从而导致下载过程中断失败。为了解决这一问题，可以由超级管理员账户登录系统，在“系统设置 > 安全策略”界面中延长“会话超时时间”，设置完成后重新登录以下载文件。具体操作步骤请参见《AI Space 管理员指南》中的“安全策略”章节。



----结束

### 3.1.3.2.6 预置数据集

组织管理员用户可以将已发布的数据集生成为预置数据集。

在数据类型为非自定义的数据集所在行，单击“更多 > 转预置”，该数据集将出现在预置数据集界面。

### 3.1.3.3 标签组管理

为了更好的在AI Space上对数据集标签进行管理，AI Space支持标签组管理。标签组分为“我的标签组”和“预置标签组”。

## 参数说明

表 3-11 参数说明

| 参数    | 说明   |
|-------|--|
| ID    | 显示标签组的ID。  |
| 名称    | 显示用户自定义的标签组的名称。  |
| 类型    | 显示标签组类型，包括： <ul style="list-style-type: none"><li>• 视觉</li><li>• 文本</li><li>• 音频</li></ul> |
| 标签数量  | 显示标签组的标签数量。  |
| 更新时间  | 显示标签组最新的编辑更新时间。  |
| 创建时间  | 显示标签组创建的时间。  |
| 标签组描述 | 显示用户自定义的标签描述。  |

### 3.1.3.3.1 我的标签组

#### 创建标签组

**步骤1** 选择“数据管理 > 标签组管理”。

进入“标签组管理”界面。

**步骤2** 选择“我的标签组”页签。

**步骤3** 单击“创建”。

进入“创建标签组”界面。

图 3-46 创建标签组

The screenshot shows the 'Create Tag Group' page. At the top, there is a back arrow and the title '创建标签组'. Below this, there are three input fields: '名称' (Name) with a placeholder '标签组名称不能超过50字' (Tag group name cannot exceed 50 characters) and a character limit of '0/50'; '类型' (Type) with a dropdown menu; and '描述' (Description) with a placeholder '标签组描述长度不能超过100字' (Tag group description length cannot exceed 100 characters) and a character limit of '0/100'. Under the '创建方式' (Creation Method) section, there are three tabs: '自定义标签组' (Custom Tag Group), '编辑标签组' (Edit Tag Group), and '导入标签组' (Import Tag Group). The '自定义标签组' tab is selected. It contains two rows for custom tags: '自定义标签1' (Custom Tag 1) with an input field '请输入标签名称' (Enter tag name) and a checkbox; and '自定义标签2' (Custom Tag 2) with an input field '请输入标签名称' (Enter tag name) and a checked checkbox. Below these fields is a note: 「自定义标签组」由用户自己创建，标签名长度不能超过 30 (Custom tag group is created by the user themselves, tag name length cannot exceed 30 characters). At the bottom is a blue '确认创建' (Confirm Creation) button.

**步骤4** 填写以下参数。

表 3-12 参数说明

| 参数 | 说明  |
|----|---|
| 名称 | 输入标签组的名称，不能超过50字符，支持中文、英文、数字、下划线和中划线。   |
| 类型 | 从下拉列表中选择标签组的类型，包括： <ul style="list-style-type: none"><li>视觉</li><li>文本</li><li>音频</li></ul> |
| 描述 | 输入标签组的描述，描述长度不能超过100字符。   |

| 参数   | 说明  |
|------|---|
| 创建方式 | <p>选择创建方式，包括：</p> <ul style="list-style-type: none"><li>• 自定义标签组：用户自己创建，输入标签名称、选择标签颜色。</li><li>• 编辑标签组：按照标准格式自由编写标签。</li><li>• 导入标签组：按照格式要求上传.json文件。</li></ul> <p>编辑标签组格式：</p> <pre>name: 名称 color: 颜色(16进制编码)</pre> <p>详细示例：</p> <pre>[{   "name": "行人",   "color": "#FFFFFF" }, {   "name": "自行车",   "color": "#000000" }]</pre> |

**步骤5** 单击“确认创建”。

----结束

## 查看标签组详情

**步骤1** 选择“数据管理 > 标签组管理”。

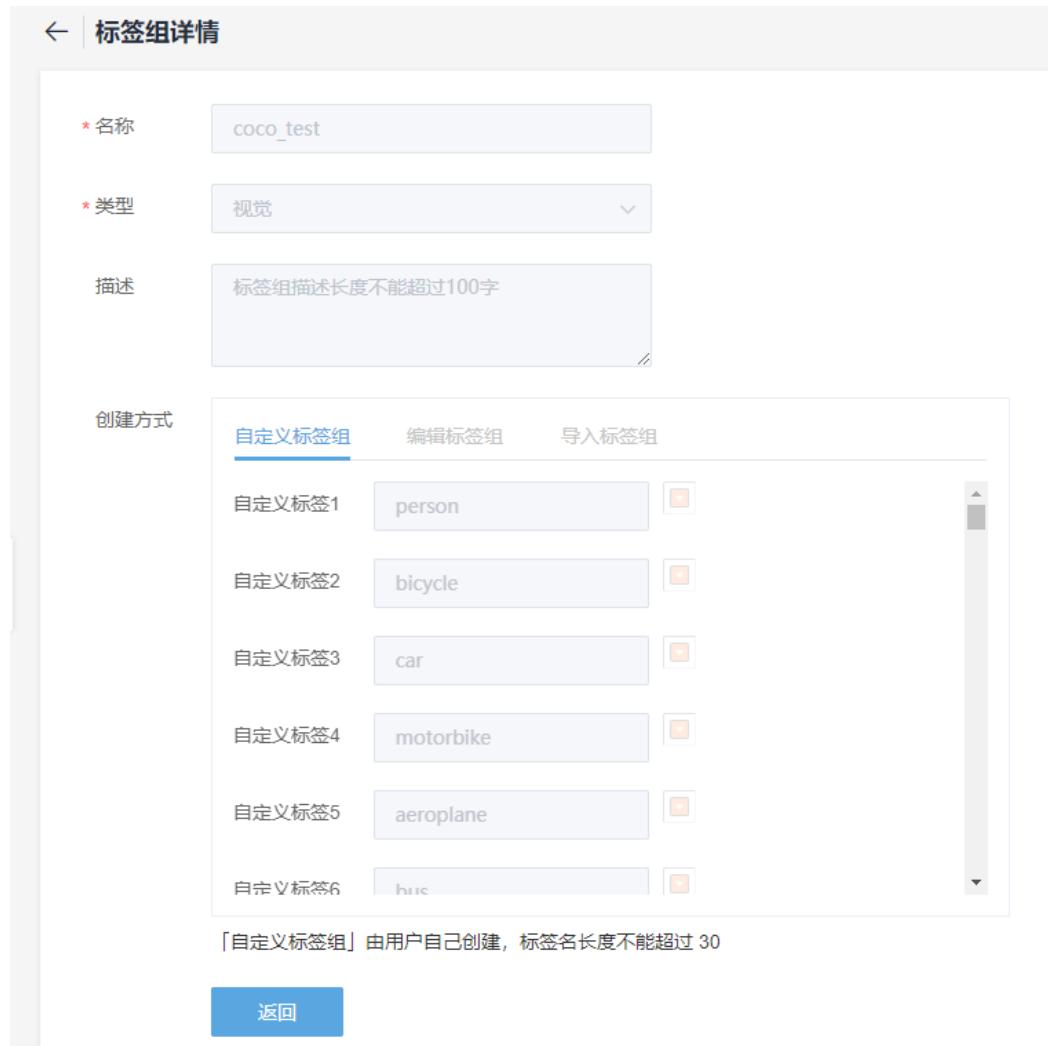
进入“标签组管理”界面。

**步骤2** 选择“我的标签组”页签。

**步骤3** 单击待查看标签组详情的标签组名称。

进入“标签组详情”页面。

图 3-47 标签组详情



----结束

## 编辑标签组

- 步骤1** 选择“数据管理 > 标签组管理”。
- 进入“标签组管理”界面。
- 步骤2** 选择“我的标签组”页签。
- 步骤3** 在待编辑的标签组所在行，单击“编辑”。
- 进入“编辑标签组”页面。

图 3-48 编辑标签组

| ID | 名称      | 类型 | 标签数量 | 更新时间                | 创建时间                | 标签组描述 | 操作   |
|----|---------|----|------|---------------------|---------------------|-------|--|
| 18 | 656     | 视觉 | 80   | 2023-05-29 19:31:26 | 2023-05-29 19:31:26 |       | <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> |
| 17 | 9632    | 视觉 | 80   | 2023-05-29 19:29:18 | 2023-05-29 19:29:18 |       | <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> |
| 16 | coco096 | 视觉 | 80   | 2023-05-29 19:27:43 | 2023-05-29 19:27:43 |       | <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> |
| 15 | coco36  | 视觉 | 80   | 2023-05-29 19:26:59 | 2023-05-29 19:26:59 |       | <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> |

步骤4 参照表3-12对标签组进行编辑。

步骤5 单击“确认编辑”。

----结束

## 复制标签组

步骤1 选择“数据管理 > 标签组管理”。

进入“标签组管理”界面。

步骤2 选择“我的标签组”页签。

步骤3 在待复制的标签组所在行，单击“复制”。

弹出“复制标签组”提示框。

图 3-49 复制标签组



**步骤4** 修改标签组的名称和描述信息。

**步骤5** 单击“确定”。

----结束

## 删除标签组

### 须知

被数据集引用的标签组不允许被删除。

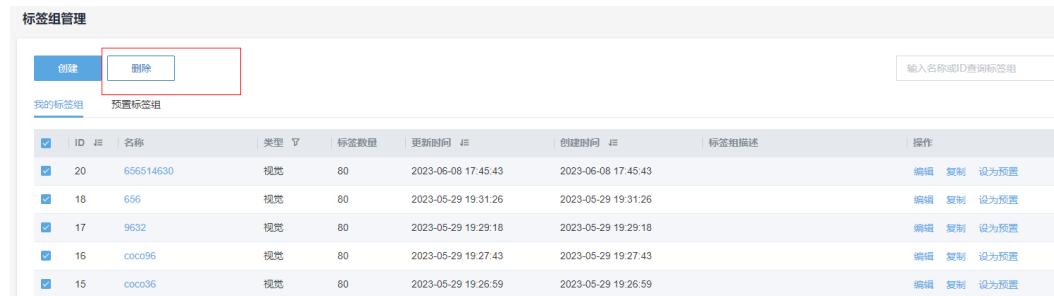
**步骤1** 选择“数据管理 > 标签组管理”。

进入“标签组管理”界面。

**步骤2** 选择“我的标签组”页签。

**步骤3** 勾选需要删除的标签组，单击“删除”。

弹出提示框。

**图 3-50 删除标签组**

| 标签组管理  |    |           |    |      |                     |                     |       |
|--|----|-----------|----|------|---------------------|---------------------|-------|
| 我的标签组  |    | 预置标签组     |    |      |                     |                     |       |
| 操作   | ID | 名称        | 类型 | 标签数量 | 更新时间                | 创建时间                | 标签组描述 |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 20 | 656514630 | 视觉 | 80   | 2023-06-08 17:45:43 | 2023-06-08 17:45:43 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 18 | 656       | 视觉 | 80   | 2023-05-29 19:31:26 | 2023-05-29 19:31:26 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 17 | 9632      | 视觉 | 80   | 2023-05-29 19:29:18 | 2023-05-29 19:29:18 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 16 | coco96    | 视觉 | 80   | 2023-05-29 19:27:43 | 2023-05-29 19:27:43 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 15 | coco36    | 视觉 | 80   | 2023-05-29 19:26:59 | 2023-05-29 19:26:59 |       |

**步骤4** 单击“确定”。

----结束

### 3.1.3.3.2 预置标签组

组织管理员用户可以将标签组转为预置标签组。

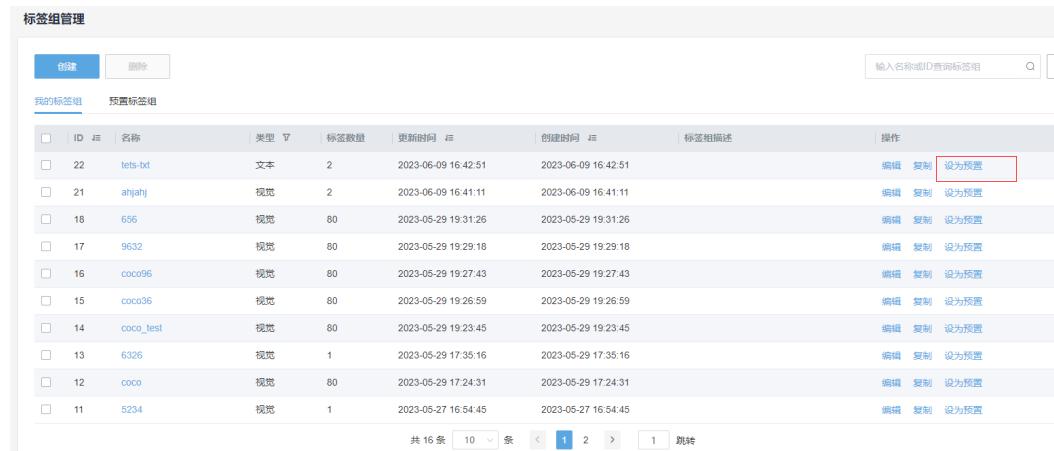
### 设为预置标签组

**步骤1** 选择“数据管理 > 标签组管理”。

进入“标签组管理”界面。

**步骤2** 选择“我的标签组”页签。

**步骤3** 在待设为预置的标签组所在行，单击“设为预置”。

**图 3-51 设为预置标签组**

| 标签组管理  |    |           |    |      |                     |                     |       |
|--|----|-----------|----|------|---------------------|---------------------|-------|
| 我的标签组  |    | 预置标签组     |    |      |                     |                     |       |
| 操作   | ID | 名称        | 类型 | 标签数量 | 更新时间                | 创建时间                | 标签组描述 |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 22 | tets-bd   | 文本 | 2    | 2023-06-09 16:42:51 | 2023-06-09 16:42:51 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 21 | ahjahj    | 视觉 | 2    | 2023-06-09 16:41:11 | 2023-06-09 16:41:11 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 18 | 656       | 视觉 | 80   | 2023-05-29 19:31:26 | 2023-05-29 19:31:26 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 17 | 9632      | 视觉 | 80   | 2023-05-29 19:29:18 | 2023-05-29 19:29:18 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 16 | coco96    | 视觉 | 80   | 2023-05-29 19:27:43 | 2023-05-29 19:27:43 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 15 | coco36    | 视觉 | 80   | 2023-05-29 19:26:59 | 2023-05-29 19:26:59 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 14 | coco_test | 视觉 | 80   | 2023-05-29 19:23:45 | 2023-05-29 19:23:45 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 13 | 6326      | 视觉 | 1    | 2023-05-29 17:35:16 | 2023-05-29 17:35:16 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 12 | coco      | 视觉 | 80   | 2023-05-29 17:24:31 | 2023-05-29 17:24:31 |       |
| <a href="#">编辑</a> <a href="#">复制</a> <a href="#">设为预置</a> | 11 | 5234      | 视觉 | 1    | 2023-05-27 16:54:45 | 2023-05-27 16:54:45 |       |

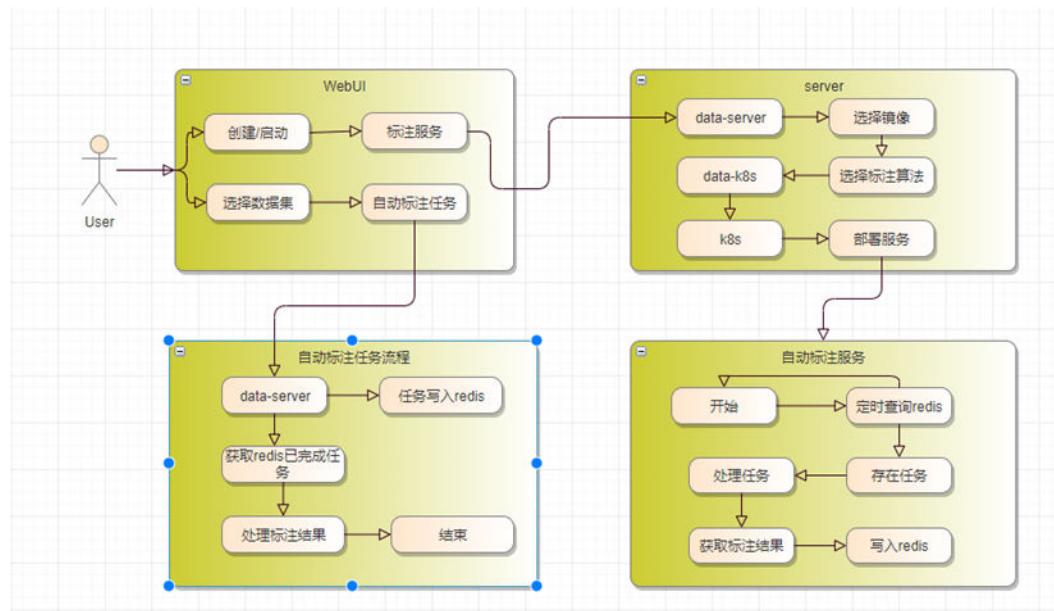
**步骤4** 该标签组将显示在预置标签组界面。

----结束

### 3.1.3.4 标注服务管理

AI Space支持图像分类，目标检测，语义分割，文本分类四种标注类型的数据集自动标注，在进行自动标注时需要选择对应的自动标注服务。在“标注服务管理”界面，用户可以通过上传算法、模型和镜像来构建自动标注服务。创建标注服务后可以查看到标注服务信息，并提供启动、停止、编辑、删除和查看标注日志等操作。

图 3-52 标注整体框架流程图



## 参数说明

表 3-13 参数说明

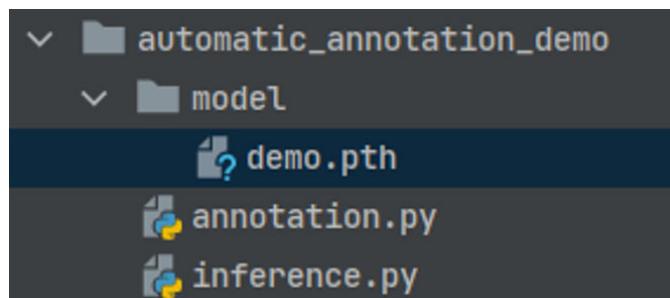
| 参数     | 说明   |
|--------|--|
| ID     | 标注服务的ID。   |
| 标注服务名称 | 标注服务的名称。   |
| 模型类型   | 标注服务的模型类型。<br>用户可单击模型类型旁的 <span style="color: #0072bc;">▼</span> ，对模型类型进行筛选。 |
| 算法名称   | 标注服务算法的名称。   |
| 镜像     | 标注服务对应的镜像。   |
| 状态     | 标注服务的状态。<br>用户可单击状态旁的 <span style="color: #0072bc;">▼</span> ，对状态进行筛选。       |
| 节点组    | 标注服务关联的节点组。  |
| 创建时间   | 标注服务创建的时间。   |
| 描述     | 标注服务的描述。   |

## 前提条件

在创建标注服务之前，需要满足以下条件：

- 选择的镜像需要有Python3环境。

- 选择的算法需按以下格式封装模型与算法，文件列表格式如图例所示。



- 若创建自动标注服务时选择了模型，则平台会创建model目录并将模型放入model目录下，若不选择模型，则忽略model目录。
- inference.py文件与annotation.py文件为平台加载该算法的必要结构。  
inference.py 文件内容如下：

```
#!/usr/bin/env python
# -*- coding:utf-8 -*-

"""
此文件必须存在，不可修改，供平台自动标注框架调用
"""

import annotation as ann

def load():
    """
    加载
    """
    print("加载")
    ann._init()

def inference(task):
    """
    推理
    """
    return ann.execute(task)
```

annotation.py文件内容如下：

```
#!/usr/bin/env python
# -*- coding:utf-8 -*-

"""
TODO 此文件为自动标注核心代码逻辑，可根据实际算法，对execute方法进行补充，入参与结果demo
已给
"""

def execute(task):
    """
    TODO 此方法必须存在,方法名与入参保持不变，入参task为json格式，与业务场景相关的字段
    为labels, files, 其余字段可不关注
    labels: 平台上数据集对应的标签名称列表
    files: 标注数据文件详情列表，id为文件标识（标注结果需带上该id），url为文件地
    址, datasetId为数据集id, name为文件名称
    task具体demo如下:
    {
        'priority': 0,
        'labels': ['cat', 'dog'],
        'files': [
            {'datasetId': 3,
             'url': '/nfs/ai-prod/tenant/pvc-1509e726-842f-4a6a-9f78-642544e243ac/dataset/3/
origin/123124.png',
             'id': 10001, 'name': '123124.png'},
            {'datasetId': 3,
             'url': '/nfs/ai-prod/tenant/pvc-1509e726-842f-4a6a-9f78-642544e243ac/dataset/3/
origin/cat.jpg',
             'id': 10002, 'name': 'cat.jpg'}
        ]
    }
```

```
'id': 10002, 'name': 'cat.jpg'}],  
'status': 1, 'labelType': 0, 'reTaskId': '3b1de129-8a65-4d31-9f71-2cae12cea49f', 'taskId': 1,  
'algorithm': 101, 'datasetId': 3, '@type': 'org.dubhe.data.domain.bo.TaskSplitBO', 'dataType': 0,  
'annotateType': 101}
```

TODO 用户可根据具体算法模型，在此方法内完善自动标注逻辑，返回结果为json格式，  
demo如下：

```
{  
    "annotations": [  
        {"annotation": [{"category_id": "cat", "bbox": [86.307, 2.655, 93.852, 129.926],  
                      "segmentation": [[130.81, 79.83], [770.51, 344.92], [615.08, 715.44], [66.47, 245.64]], "score":  
                      0.96}], "id": 10001},  
        {"annotation": [{"category_id": "cat", "bbox": [86.307, 2.655, 93.852, 129.926],  
                      "segmentation": [[130.81, 79.83], [770.51, 344.92], [615.08, 715.44], [66.47, 245.64]], "score":  
                      0.96}], "id": 10002}  
    ]  
}
```

字段解释：

整体格式为 {"annotations" : 标注结果数组 }

标注结果数组：[标注结果1, 标注结果2 ...]

标注结果： {"annotations" : 标注结果json数组字符串, "id" : task传入的文件标识 }

标注结果json数组字符串：[标注情况1, 标注情况2 ...]

标注情况： {

```
    "category_id": "cat",  
    "score": 0.96,  
    "bbox": [86.307, 2.655, 93.852, 129.926],  
    "segmentation": [[130.81, 79.83], [770.51, 344.92], [615.08, 715.44], [66.47, 245.64]]
```

}

category\_id: 名称固定, value为task中传入的标签名称

score: 标注评分

bbox: 目标检测坐标

segmentation: 语义分割坐标

...

# TODO 根据算法模型，完成相关标注逻辑，按格式返回结果，如下为目标检测的返回demo

```
result_demo = {'annotations': [  
    {  
        'annotation': [{"category_id": "cat", "score": 0.99, "bbox": [86.30, 2.62, 93.85, 29.92]}],  
        'id': 10001},  
    {  
        'annotation': [{"category_id": "cat", "score": 0.96, "bbox": [94.58, 17.87, 53.26, 74.86]}],  
        'id': 10002}  
]
```

```
return result_demo
```

```
def __init__:
```

TODO 此方法必须存在，为初始化加载算法或者模型，并将全局处理

例子如下，若用户场景无需将模型提前加载至内存，也可不加多余代码，但方法必须有

""

```
print('init yolo_obj')  
# global yolo_obj  
# yolo_obj = yolo_demo.YoloInference()
```

- 参考[上传算法、上传模型、3.1.8.2 创建镜像](#)章节将准备好的可用算法、模型和镜像上传至平台。

## 创建标注服务

**步骤1** 选择“数据管理 > 标注服务管理”。

进入“标注服务管理”界面。

**步骤2** 单击“创建”。

进入“创建服务”页面。

图 3-53 创建服务

创建服务

\* 服务名称: 服务名称不能超过32字

\* 模型类型: 请选择

\* 算法: 请选择

预加载模型: 请选择模型 请选择模型版本

\* 镜像: 请选择镜像 请选择镜像版本

\* 节点组: 请选择

\* 计算类型: 基础资源

\* 基础资源: 2CPU 4GBMEM

CPU: 2核

内存: 4096Mi

\* 服务数量: 1

调度器类型: kube-scheduler

描述: 服务描述

取消 确定

**步骤3** 填写如下参数。

表 3-14 参数说明

| 参数    | 说明  |
|-------|---|
| 服务名称  | 输入标注服务名称，支持字母、数字、汉字、英文横杠和下划线，名称长度不能超过32字。 |
| 模型类型  | 从下拉列表中选择模型类型。                             |
| 算法    | 从下拉列表中选择准备好的算法。                           |
| 预加载模型 | 从下拉列表中选择预加载模型和模型版本。                       |
| 镜像    | 从下拉列表中选择上传好的镜像和镜像版本。                      |

| 参数        | 说明  |
|-----------|---|
| 节点组       | 从下拉列表中选择该标注服务关联的节点组。  |
| 计算类型      | 根据不同的算法选择标注服务需要的计算节点类型，包括：<br><ul style="list-style-type: none"> <li>● 基础资源</li> <li>● GPU / 整卡</li> <li>● GPU / MIG</li> <li>● NPU / 整卡</li> <li>● NPU / vNPU</li> </ul> |
| 基础资源      | 从下拉框选择基础资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。<br>选择完成后，将显示该基础资源规格参数信息，包括CPU个数、内存大小。  |
| GPU/NPU规格 | （当计算类型为GPU或NPU时需要选择）从下拉列表中选择资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。<br>选择完成后，将显示该资源规格参数信息。   |
| 服务数量      | 选择服务数量，默认为1。  |
| 描述        | 输入标注服务的描述。  |

表 3-15 调度器配置

| 参数       | 说明   |
|----------|--|
| 调度器类型    | AI Space的调度器类型，超级管理员用户可在控制台 > 调度配置界面修改。<br>当调度器类型为Volcano时，填写下方调度参数。   |
| 所属队列     | 从下拉列表中选择标注服务任务所属队列。  |
| 优先级      | 选择标注服务任务的优先级。<br>当集群中运行多个job时，Volcano将以用户定义的优先级调度资源。   |
| 最少运行pod数 | 设置标注服务任务最少运行pod数，最少运行pod数需小于节点数。   |
| 最大重启次数   | 设置标注服务任务最大重启次数。  |
| 生命周期策略   | 选择Pod生命周期策略。支持“pod驱逐重启作业”，当pod被驱逐时，将重启该job。  |
| 节点亲和性    | 输入节点属性，即节点标签，设置节点亲和性前需要为节点打上对应的标签。<br>单击  或  可添加或删除属性。 |

| 参数 |     | 说明  |
|----|-----|---|
|    | 类型  | <p>从下拉列表中选择节点亲和性类型，包括硬亲和性调度策略、软亲和性调度策略。</p> <ul style="list-style-type: none"> <li>硬亲和性调度策略：调度器必须满足，多条规则间是一种“或”的关系，即只需要满足一条规则即会进行调度。</li> <li>软亲和性调度策略：调度器会尽量满足，无论是满足其中一条或者是都不满足都会进行调度。</li> </ul>                            |
|    | 操作符 | <p>从下拉列表中选择操作符。</p> <ul style="list-style-type: none"> <li>In：标签的值在某个列表中</li> <li>NotIn：标签的值不在某个列表中</li> <li>Exists：某个标签存在</li> <li>DoesNotExist：某个标签不存在</li> <li>Gt：标签的值大于某个值（字符串比较）</li> <li>Lt：标签的值小于某个值（字符串比较）</li> </ul> |
|    | 取值  | <p>添加属性取值。<br/>通过配置节点亲和性规则，调度器可以将Pod调度到具有特定标签的节点。</p>   |

**步骤4** 单击“确定”，列表页提示创建成功。

新建的标注服务默认启动，启动完成后为“运行中”状态。

#### 说明

出现启动失败可能的原因如下：

- 配置的节点数或规格资源不足
- 参数配置、模型错误
- 模型与算法脚本不匹配
- 镜像错误等

出现启动失败请查看日志信息。

----结束

## 停止标注服务

**步骤1** 选择“数据管理 > 标注服务管理”。

进入“标注服务管理”界面。

**步骤2** 对于状态为“运行中”的标注服务，单击“停止”。

弹出提示框。

图 3-54 停止标注服务



步骤3 单击“确定”，停止标注服务。

----结束

## 修改标注服务

步骤1 选择“数据管理 > 标注服务管理”。

进入“标注服务管理”界面。

步骤2 对于状态为“已停止”的标注服务，单击操作列的“编辑”。

进入编辑界面。

步骤3 根据需要参考表3-14修改信息。

步骤4 单击“确定”，完成修改。

----结束

## 启动标注服务

步骤1 选择“数据管理 > 标注服务管理”。

进入“标注服务管理”界面。

步骤2 对于状态为“已停止”的标注服务，单击操作列的“启动”。

状态从“启动中”变为“运行中”状态。

图 3-55 启动服务



## 说明

出现启动失败可能的原因如下：

- 配置的节点数或规格资源不足
- 参数配置、模型错误
- 模型与算法脚本不匹配
- 镜像错误等

出现启动失败请查看日志信息。

----结束

## 查看日志

**步骤1** 选择“数据管理 > 标注服务管理”。

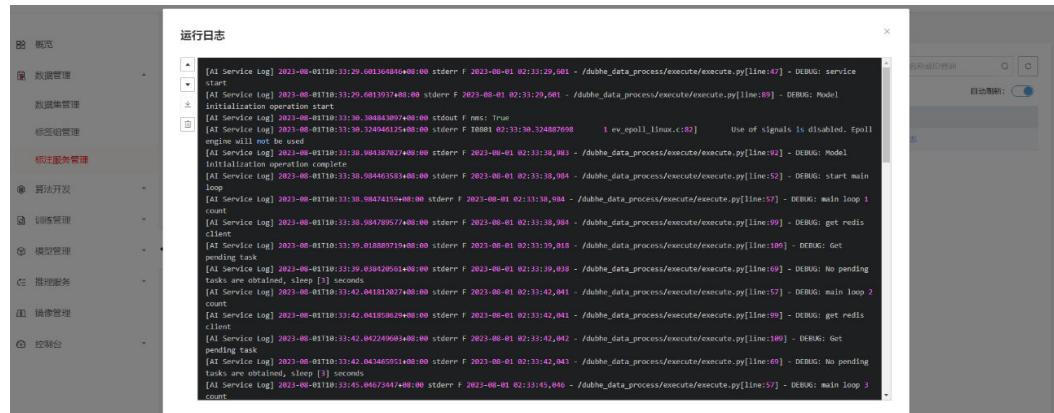
进入“标注服务管理”界面。

**步骤2** 单击待查看的标注服务所在行的“日志”。

弹出“运行日志”提示框。

**步骤3** 查看当前运行中的标注服务信息。

图 3-56 查看日志



----结束

## 删除标注服务

**步骤1** 选择“数据管理 > 标注服务管理”。

进入“标注服务管理”界面。

**步骤2** 勾选状态为“已停止”或“运行失败”的标注服务。

**步骤3** 单击界面左上角的“删除”。

弹出提示框。

**步骤4** 单击“确定”，删除该标注服务。

----结束

## 3.1.4 算法开发

### 3.1.4.1 算法开发简介

算法开发模块是AI Space提供给用户开发算法的模块，包含Notebook和算法管理两部分：

- Notebook是一种交互式编程环境，AI开发者可以在云端进行机器学习的开发。该模块集成了开源的JupyterLab，可支持开发者在线编辑、调试、运行代码，同时预置了PyTorch，TensorFlow等多种深度学习框架，用户可在多种框架之间自由切换。每个Notebook都是一个独立的编程环境，用户可以对Notebook实现创建、打开、停止、启动、删除等操作。算法开发完成之后，还可以将其保存到“算法管理”的“我的算法”中，进行后续的训练工作。
- 算法管理用于保存一些AI Space预置的算法以及用户自己开发的算法。对于已经保存的算法，用户可以进行在线编辑、创建训练任务、下载、Fork、删除等操作。

### 3.1.4.2 Notebook

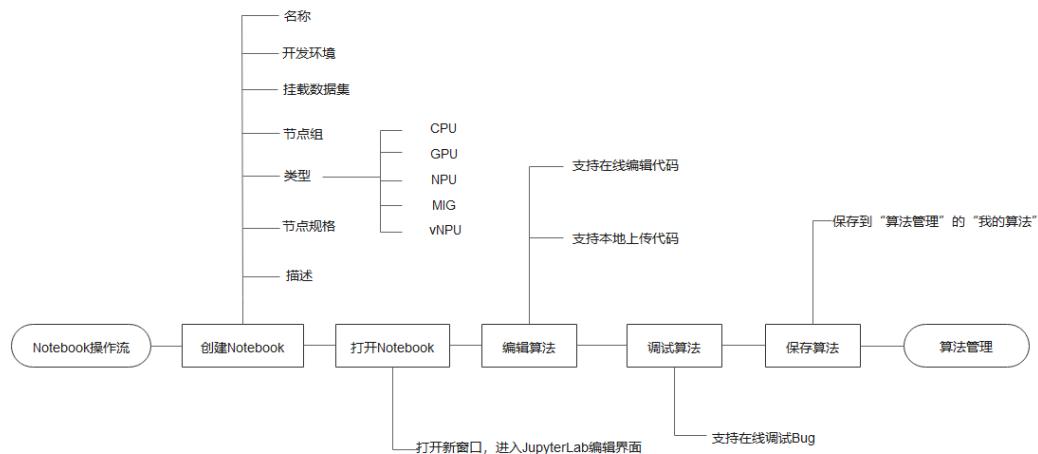
用户可在该界面中查询到属于自己创建的Notebook信息，并可对Notebook进行创建、打开、停止、启动、删除、保存算法、搜索等操作。

#### 须知

Notebook默认将会在启动后八小时自动关闭，请及时保存您的代码。

## Notebook 操作流程

图 3-57 算法开发流程图



- 当用户需要从无到有开发算法时，首先需要创建Notebook，期间可以根据自己的需要进行一系列相关配置。
- Notebook创建完成后，即可在界面上单击“打开”进入JupyterLab界面进行代码编辑。如果用户在本地已有代码，可直接上传。

- 代码编辑完成后，可进行在线运行代码。
- 算法开发完成后，可单击“保存算法”，将开发完成的算法保存到“算法管理”的“我的算法”中。

## 参数介绍

表 3-16 参数说明

| 参数   | 说明                   |
|------|----------------------|
| 名称   | 显示用户自定义的Notebook的名称。 |
| CPU  | 显示Notebook使用的CPU数量。  |
| 内存   | 显示Notebook使用的内存大小。   |
| 计算类型 | 显示Notebook的计算类型。     |
| 资源数量 | 显示Notebook的资源数量。     |
| 描述   | 显示用户自定义的Notebook的描述。 |
| 状态   | 显示Notebook的状态。       |
| 节点组  | 显示Notebook关联的节点组。    |
| 创建时间 | 显示Notebook创建的时间。     |

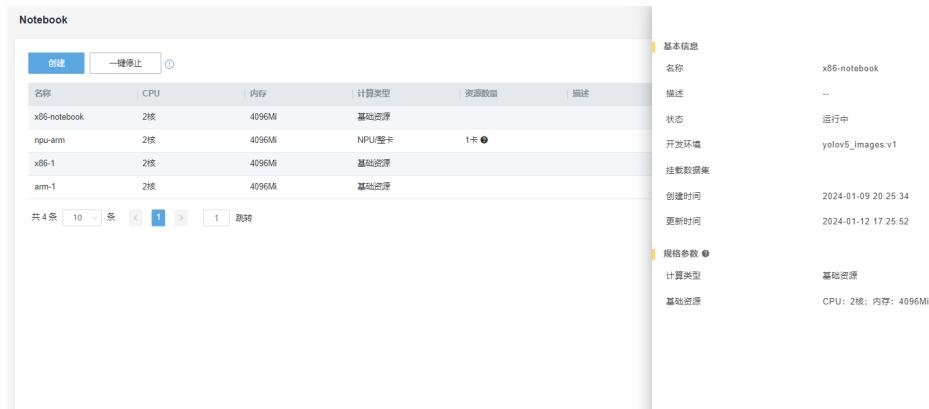
## 查看 Notebook 详情

步骤1 选择“算法开发 > Notebook”。

进入“Notebook”页面。

步骤2 单击列表中任意一条Notebook，在右侧弹窗中可查看具体的基本信息和规格参数。

图 3-58 Notebook 详情



----结束

## 创建 Notebook

### 说明

- Notebook镜像制作打包过程请参见[3.1.4.2.1 Notebook镜像打包指导](#)。
- Notebook镜像制作完毕后，需上传至AI Space后才能使用，请参考[3.1.8 镜像管理](#)章节上传镜像。

**步骤1** 选择“算法开发 > Notebook”。

进入“Notebook”页面。

**步骤2** 单击“创建”。

弹出“创建Notebook”提示框。

**图 3-59** 创建 Notebook

The screenshot shows the 'Create Notebook' dialog box. It includes fields for Name, Development Environment, Data Set, Node Group, Compute Type, and Basic Resources. It also displays CPU and Memory specifications, a Description area, and two buttons at the bottom: 'Cancel' and 'Confirm'.

| 参数   | 值                    |
|------|----------------------|
| 名称   | 请输入 Notebook 名称 0/30 |
| 开发环境 | 请选择镜像 请选择镜像版本        |
| 数据集  | 请选择数据集 请选择数据集版本      |
| 节点组  | 请选择                  |
| 计算类型 | 基础资源                 |
| 基础资源 | 2CPU 4GBMEM          |
| CPU  | 2核                   |
| 内存   | 4096Mi               |
| 描述   | 0/255                |

**步骤3** 填写如下参数。

表 3-17 参数说明

| 参数        | 说明  |
|-----------|---|
| 名称        | 输入Notebook名称，支持字母、数字、汉字、英文横杠和下划线，且英文横杠不可在首尾。  |
| 开发环境      | 选择镜像及镜像版本。  |
| 数据集       | 选择数据集及数据集版本。  |
| 节点组       | 从下拉列表中选择该Notebook关联的节点组。  |
| 计算类型      | 选择Notebook需要的计算类型，包括： <ul style="list-style-type: none"><li>● 基础资源</li><li>● GPU / 整卡</li><li>● GPU / MIG</li><li>● NPU / 整卡</li><li>● NPU / vNPU</li></ul> |
| 基础资源      | 从下拉框选择基础资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。<br>选择完成后，将显示该基础资源规格参数信息，包括CPU个数、内存大小。  |
| GPU/NPU规格 | （当计算类型为GPU或NPU时需要选择）从下拉列表中选择资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。<br>选择完成后，将显示该资源规格参数信息。   |
| 描述        | 输入Notebook描述。   |

**步骤4** 单击“确定”，Notebook列表页提示创建成功。

新建的Notebook默认启动。

启动完成后为“运行中”状态，可进行打开、停止、保存算法等操作。

----结束

## 打开 Notebook

**步骤1** 选择“算法开发 > Notebook”。

进入“Notebook”页面。

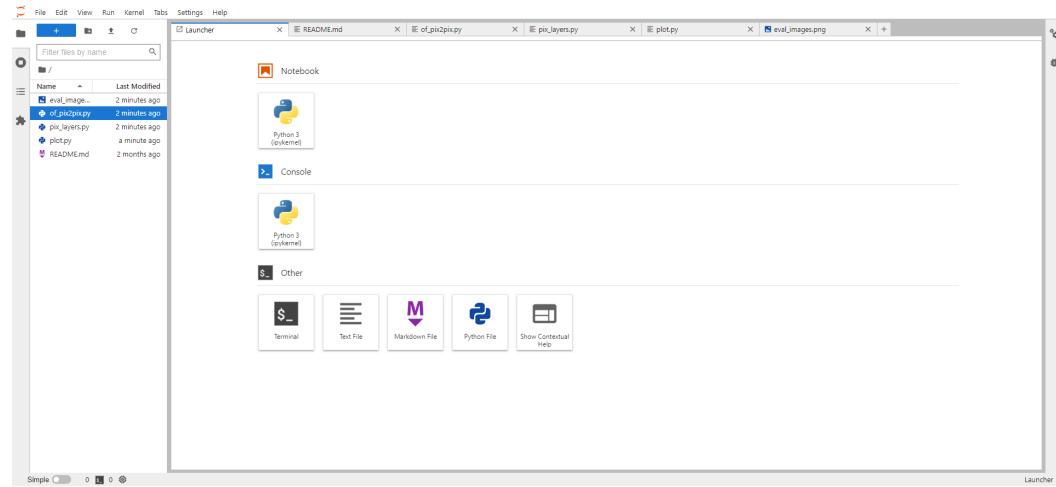
**步骤2** 选择运行状态的Notebook，单击“打开”。

进入相应的Notebook的编辑界面。

**步骤3** 进行算法文件的创建和编辑。

在左侧进行文件夹/文件的新建与导入，在右侧选择运行程序的环境，目前支持Python3，PyTorch，TensorFlow等。

图 3-60 JupyterLab 创建页面



可在此开发环境中编写、调试、保存代码程序：

图 3-61 JupyterLab 编辑页面

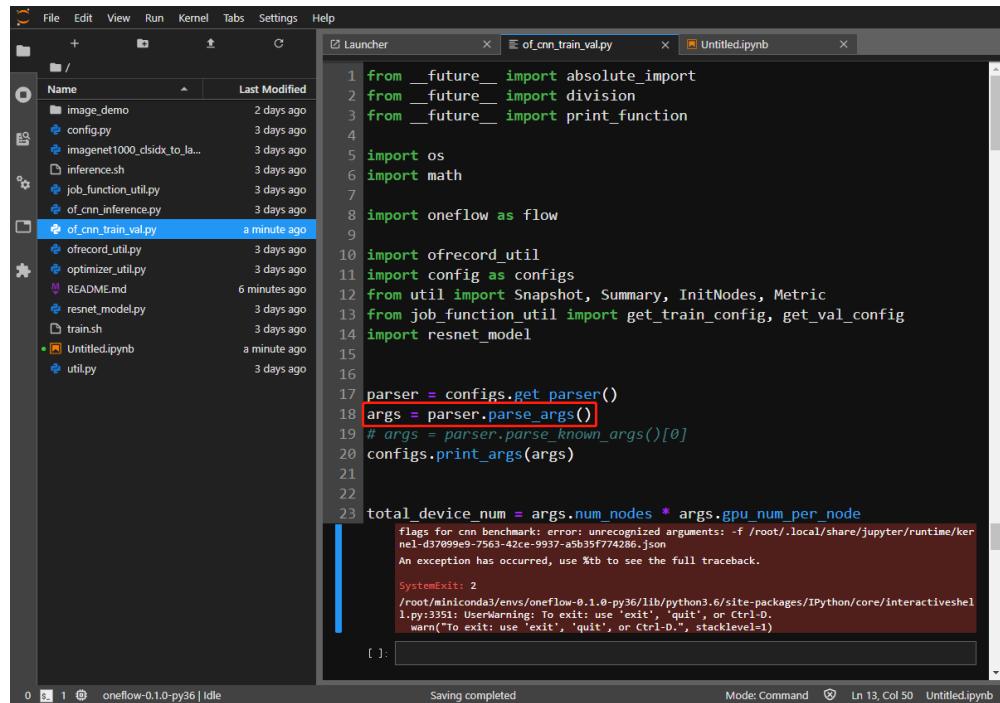
```
1  #!/usr/bin/env python
2  # Copyright 2010 The OneFlow Authors. All rights reserved.
3  # Licensed under the Apache License, Version 2.0 (the "License");
4  # you may not use this file except in compliance with the License.
5  # You may obtain a copy of the License at
6  # http://www.apache.org/licenses/LICENSE-2.0
7  # Unless required by applicable law or agreed to in writing, software
8  # distributed under the License is distributed on an "AS IS" BASIS,
9  # WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
10 # See the License for the specific language governing permissions and
11 # limitations under the License.
12 #
13
14 import oneflow as flow
15 from typing import Tuple
16 import oneflow.typing as tp
17 import os
18 import imghdr
19 import os
20 import pix2pix as layers
21 import matplotlib.pyplot as plt
22 import time
23 import torch
24 os.environ["CUDA_VISIBLE_DEVICES"] = '1'
25
26 class Pix2Pix:
27     def __init__(self, args):
28         self.args = args.learning_rate
29         self.image_size = 256
30         self.image_mean = 0.5
31         self.image_std = 1.0
32         self.label_smooth = args.label_smooth
33
34         self.batch_size = args.batch_size
35         self.path = args.path
36         if not os.path.exists(self.path):
37             os.mkdir(self.path)
38             print("new dir [{}]. done - {}".format(self.path))
39         self.checkpoint_path = os.path.join(self.path, "checkpoint")
40         if not os.path.exists(self.checkpoint_path):
41             os.mkdir(self.checkpoint_path)
42         self.test_image_path = os.path.join(self.path, "test_images")
43         if not os.path.exists(self.test_image_path):
44             os.mkdir(self.test_image_path)
```

----结束

## Notebook 注意事项

Notebook不支持parse\_args()函数，如果程序当中用到这个函数，Notebook执行时会报错。

图 3-62 执行 parse\_args() 报错



The screenshot shows a Jupyter Notebook interface. On the left is a file tree with several Python files and scripts. The central area contains the code for `of_cnn_train_val.py`:

```
1 from __future__ import absolute_import
2 from __future__ import division
3 from __future__ import print_function
4
5 import os
6 import math
7
8 import oneflow as flow
9
10 import ofrecord_util
11 import config as configs
12 from util import Snapshot, Summary, InitNodes, Metric
13 from job_function_util import get_train_config, get_val_config
14 import resnet_model
15
16
17 parser = configs.get_parser()
18 args = parser.parse_args()
19 # args = parser.parse_known_args()[0]
20 configs.print_args(args)
21
22
23 total_device_num = args.num_nodes * args.gpu_num_per_node
flags for cnn_benchmark: error: unrecognized arguments: -f /root/.local/share/jupyter/runtime/kernel-d37099e9-7563-42ce-9937-a5b35f774286.json
An exception has occurred, use %tb to see the full traceback.

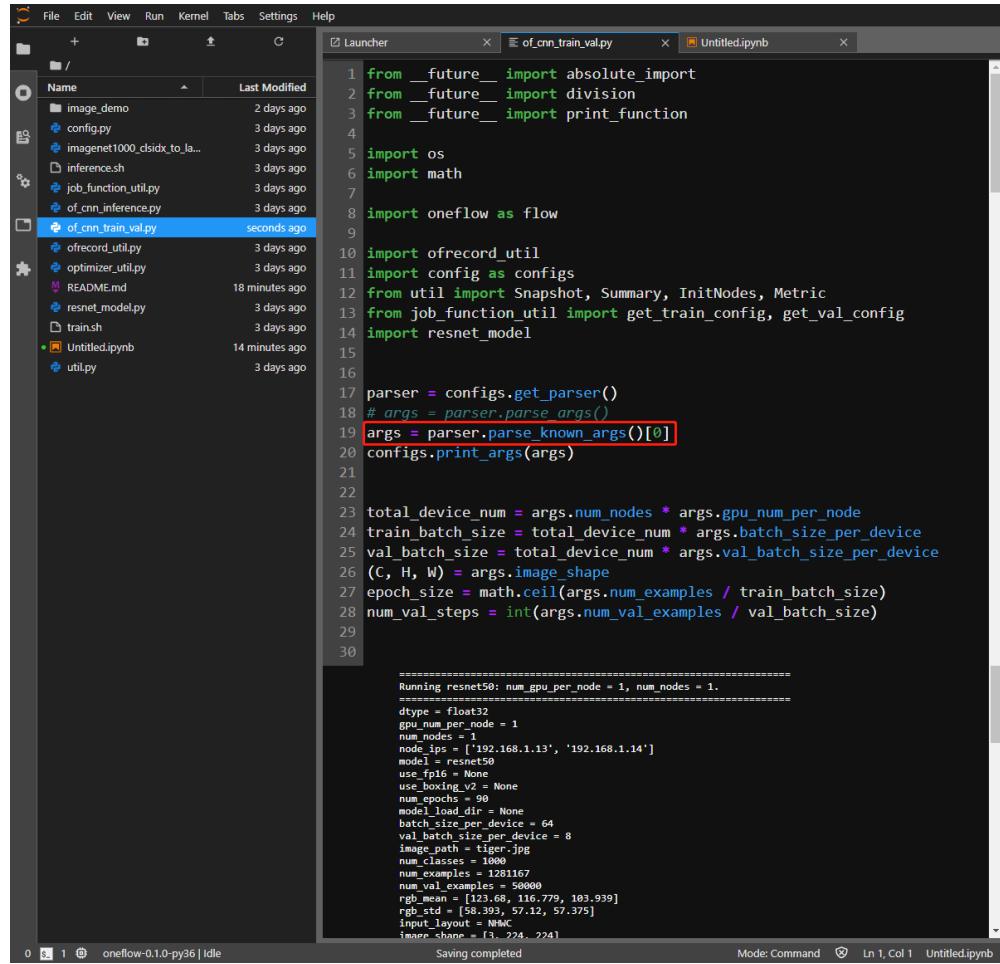
SystemExit: 2
/root/miniconda3/envs/oneflow-0.1.0-py36/lib/python3.6/site-packages/IPython/core/interactiveshell.py:3351: UserWarning: To exit: use 'exit', 'quit', or Ctrl-D.
warn("To exit: use 'exit', 'quit', or Ctrl-D.", stacklevel=1)
```

The line `args = parser.parse_args()` is highlighted with a red box. The output pane at the bottom shows a warning message:

flags for cnn\_benchmark: error: unrecognized arguments: -f /root/.local/share/jupyter/runtime/kernel-d37099e9-7563-42ce-9937-a5b35f774286.json  
An exception has occurred, use %tb to see the full traceback.  
SystemExit: 2  
/root/miniconda3/envs/oneflow-0.1.0-py36/lib/python3.6/site-packages/IPython/core/interactiveshell.py:3351: UserWarning: To exit: use 'exit', 'quit', or Ctrl-D.  
warn("To exit: use 'exit', 'quit', or Ctrl-D.", stacklevel=1)

将`parse_args()`函数改为`parse_known_args()[0]`函数即可解决。

图 3-63 解决方法



The screenshot shows a Jupyter Notebook interface with a file tree on the left and a code editor on the right. The code editor displays Python code for training a CNN. A specific line of code, `args = parser.parse\_known\_args()[0]`, is highlighted with a red box. The output pane below the code editor shows the execution results, including configuration parameters like `total\_device\_num` and `train\_batch\_size`. The status bar at the bottom indicates the mode is 'Command'.

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

import os
import math

import oneflow as flow

import ofrecord_util
import config as configs
from util import Snapshot, Summary, InitNodes, Metric
from job_function_util import get_train_config, get_val_config
import resnet_model

parser = configs.get_parser()
# args = parser.parse_args()
args = parser.parse_known_args()[0]
configs.print_args(args)

total_device_num = args.num_nodes * args.gpu_num_per_node
train_batch_size = total_device_num * args.batch_size_per_device
val_batch_size = total_device_num * args.val_batch_size_per_device
(C, H, W) = args.image_shape
epoch_size = math.ceil(args.num_examples / train_batch_size)
num_val_steps = int(args.num_val_examples / val_batch_size)

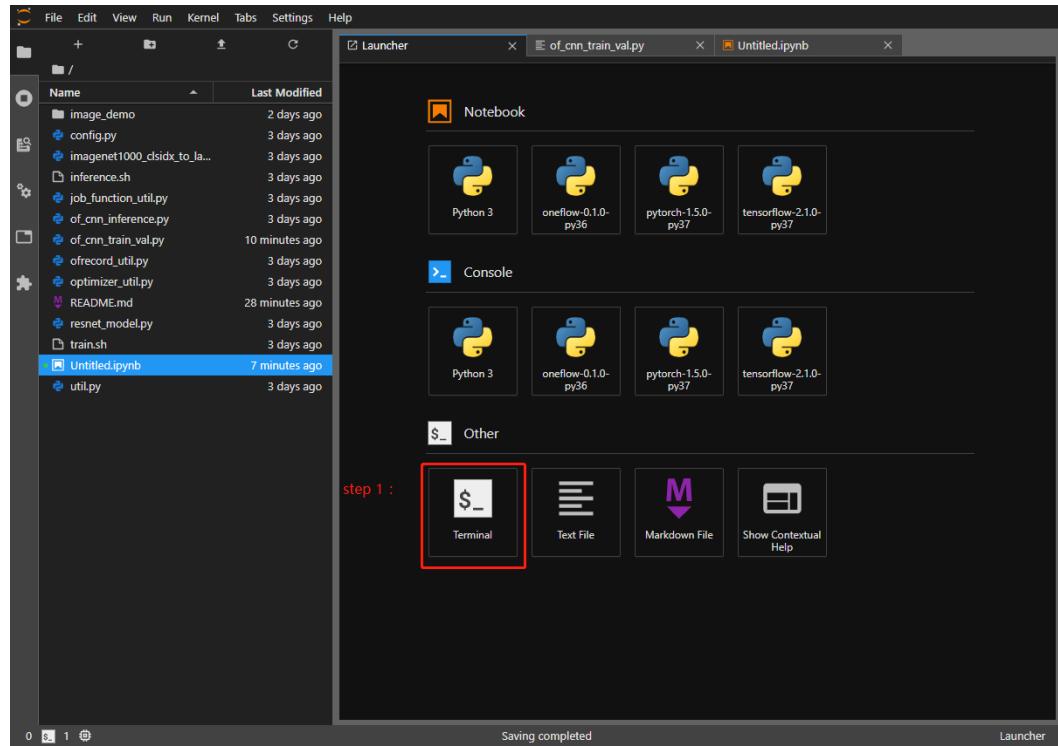
=====
Running resnet50: num_gpu_per_node = 1, num_nodes = 1.
=====
dtype = float32
gpu_num_per_node = 1
num_nodes = 1
node_ip = ['192.168.1.13', '192.168.1.14']
model = resnet50
use_fp16 = None
use_mean_image_norm = None
num_epochs = 90
model_load_dir = None
batch_size_per_device = 64
val_batch_size_per_device = 8
image_path = tiger.jpg
num_classes = 1000
num_examples = 1281167
num_val_examples = 50000
rgb_mean = [123.68, 116.779, 103.939]
rgb_std = [58.393, 57.12, 57.375]
input_layout = NHWC
image_shape = [3, 224, 224]
```

## 安装所需软件包操作步骤

用户在开发算法的过程中用到的软件包是多种多样的，Notebook当中提供了一种方法可以让用户安装自己所需要的软件包。

**步骤1 打开终端。**

图 3-64 打开终端



**步骤2** 启动bash。

**步骤3** 启动相应虚拟环境。

**步骤4** 安装所需软件包，这里以安装SciPy为例。

安装完成后即可使用SciPy包进行开发。

图 3-65 安装软件包

```
root@notebook-rn-2020071: ~
# ./bin/bash
step 2
(base) root@notebook-rn-2020071014231356610lu-lym4j-0:/# conda env list
# conda environments:
#
base          * /root/miniconda3
oneflow-0.1.0-py36      /root/miniconda3/envs/oneflow-0.1.0-py36
python3           /root/miniconda3/envs/python3
pytorch-1.5.0-py37      /root/miniconda3/envs/pytorch-1.5.0-py37
tensorflow-2.1.0-py37     /root/miniconda3/envs/tensorflow-2.1.0-py37

(base) root@notebook-rn-2020071014231356610lu-lym4j-0:/# conda activate oneflow-0.1.0-py36
step 3
(oneflow-0.1.0-py36) root@notebook-rn-2020071014231356610lu-lym4j-0:/#
(oneflow-0.1.0-py36) root@notebook-rn-2020071014231356610lu-lym4j-0:/#
(oneflow-0.1.0-py36) root@notebook-rn-2020071014231356610lu-lym4j-0:/#
(oneflow-0.1.0-py36) root@notebook-rn-2020071014231356610lu-lym4j-0:/#
(oneflow-0.1.0-py36) root@notebook-rn-2020071014231356610lu-lym4j-0:/# pip install scipy
step 4
Collecting scipy
  Downloading scipy-1.5.1-cp36-cp36m-manylinux1_x86_64.whl (25.9 MB)
    |
    |████████████████████████████████| 25.9 MB 4.1 MB/s
Requirement already satisfied: numpy>=1.14.5 in /root/miniconda3/envs/oneflow-0.1.0-py36/lib/python3.6/site-packages (from scipy) (1.19.0)
Installing collected packages: scipy
Successfully installed scipy-1.5.1
(oneflow-0.1.0-py36) root@notebook-rn-2020071014231356610lu-lym4j-0:/#
```

----结束

### 3.1.4.2.1 Notebook 镜像打包指导

#### 2.1. Nvidia 镜像与普通镜像处理步骤

##### 适用场景

- 若该Notebook需要调用GPU，则需要以dockerhub上的nvidia/cuda镜像为基础镜像。因为Ubuntu镜像建的容器无法调用显卡，即使手动安装显卡驱动后，也无法调用显卡，所以需要使用nvidia/cuda作为基础镜像。
- 若该Notebook无需调用GPU，则普通Ubuntu镜像（镜像执行用户必须是root）等均需按照如下步骤处理。

本章节使用镜像nvaeie\_chat来演示镜像打包的流程。

##### 说明

以下操作需要在有docker环境下才能够执行。若无docker，则不支持以下操作。

#### 方案一：使用 Dockerfile 文件

Dockerfile示例：

```
FROM nvidia/cuda:12.0.0-devel-ubuntu20.04
# 注意，该Dockerfile为NVIDIA样例，请按实际情况修改，具体步骤在方案二体现
# 安装Python 与 Jupyterlab
RUN apt-get update && apt-get install -y python3-dev && \
    apt-get install -y python3-pip && python3 -m pip install jupyter && \
    python3 -m pip install ipywidgets && jupyter nbextension enable --py widgetsnbextension && \
    python3 -m pip install jupyterlab && python3 -m pip install jupyterlab
# 修改Jupyter 配置文件
RUN mkdir /workspace && jupyter notebook --generate-config && \
    sed -i "s|c.NotebookApp.notebook_dir = '|c.NotebookApp.notebook_dir = '/workspace'"g" /root/.jupyter/
jupyter_notebook_config.py && \
    echo "touch /workspace/README.md" >> /etc/profile && \
    echo "echo "如需使用本算法进行训练任务，需在要执行的Python文件中接收如下参数：" >> /etc/profile && \
    echo "data_url: 数据集路径" >> /etc/profile && \
    echo "data_url: 数据集路径" >> /etc/profile && \
    echo "train_log: 日志输出" >> /etc/profile && \
    echo "训练时这些参数名会以命令行传参的形式传入可执行的Python文件中，而参数值会由系统指定。例如：" >> /etc/profile && \
    echo "python run.py --key1=value1 .... --data_url=/data/ --train_out=/out/ --train_log=/log/" >> /etc/profile && \
    echo "所以如果可执行的Python文件不接收这些参数值，可能会导致训练任务失败。" > /workspace/
README.md' >> /etc/profile
WORKDIR /workspace
CMD /bin/bash -c "source /etc/profile && jupyter lab --ip=0.0.0.0 --port=8888 --no-browser --allow-root"
```

#### 方案二：docker commit 命令

步骤1 用此镜像起容器。

```
docker run --runtime=nvidia --name=jupyterlab_v1 -dt nvaeie_chat
```

```
[root@node02 ~]# docker run --runtime=nvidia --name=jupyter_v1 -dt nvaeie_chat
48760329c039ca67f9e8974f9edee2ac19ef09e26e4afaef8ba26247bd9ec548
```

步骤2 查看容器。

```
docker ps | grep jupyterlab
```

进入容器。

```
docker exec -it cad9 bash
```

```
[root@node02 ~]# docker ps |grep jupyter_v1
48760329c039      nivate_chat      "/opt/nvidia/nvidia_--"   23 seconds ago      Up 22 seconds      6006/tcp,
jupyter_v1
[root@node02 ~]#
[root@node02 ~]#
[root@node02 ~]# docker exec -it 4876 bash
root@48760329c039:/workspace#
```

### 步骤3 在容器中，安装Jupyterlab。

#### 1. 更新软件列表。

```
apt-get update
```

```
root@48760329c039:/workspace# apt-get update
Get:1 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Hit:2 http://archive.ubuntu.com/ubuntu focal InRelease
Get:3 http://archive.ubuntu.com/ubuntu focal-updates InRelease [114 kB]
Get:4 http://security.ubuntu.com/ubuntu focal-security/universe amd64 Packages [1056 kB]
Get:5 http://archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
Get:6 http://archive.ubuntu.com/ubuntu focal-updates/universe amd64 Packages [1351 kB]
Get:7 http://security.ubuntu.com/ubuntu focal-security/main amd64 Packages [2773 kB]
Get:8 http://archive.ubuntu.com/ubuntu focal-updates/main amd64 Packages [3252 kB]
Get:9 http://security.ubuntu.com/ubuntu focal-security/restricted amd64 Packages [2345 kB]
Get:10 http://archive.ubuntu.com/ubuntu focal-updates/restricted amd64 Packages [2483 kB]
Get:11 http://archive.ubuntu.com/ubuntu focal-backports/universe amd64 Packages [28.6 kB]
Get:12 http://archive.ubuntu.com/ubuntu focal-backports/main amd64 Packages [55.2 kB]
Fetched 13.7 MB in 6s (2242 kB/s)
Reading package lists... Done
```

#### 2. 安装Python。

```
apt-get install -y python3-dev
```

```
root@48760329c039:/workspace# apt-get install -y python3-dev
Reading package lists... Done
Building dependency tree
Reading state information... Done
python3-dev is already the newest version (3.8.2-0ubuntu2).
0 upgraded, 0 newly installed, 0 to remove and 66 not upgraded.
```

#### 3. 安装pip。

```
apt-get install -y python3-pip
```

```
root@48760329c039:/workspace# apt-get install -y python3-pip
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  python-pip-whl python3-pkg-resources python3-setuptools python3-wheel
Suggested packages:
  python-setuptools-doc
The following NEW packages will be installed:
  python-pip-whl python3-pip python3-pkg-resources python3-setuptools python3-wheel
0 upgraded, 5 newly installed, 0 to remove and 66 not upgraded.
```

#### 4. 安装jupyter。

```
python3 -m pip install jupyter
```

```
root@48760329c039:/workspace# python3 -m pip install jupyter -i http://mirrors.aliyun.com/pypi/simple --trusted-host m
aliyun.com
Looking in indexes: http://mirrors.aliyun.com/pypi/simple, https://pypi.ngc.nvidia.com
Collecting jupyter
  Downloading http://mirrors.aliyun.com/pypi/packages/83/df/0f5dd132200728a86190397e1ea87cd76244e42d39ec5e88efd25b2abd
er-1.0.0-py2.py3-none-any.whl (2.7 kB)
Requirement already satisfied: notebook in /usr/local/lib/python3.8/dist-packages (from jupyter) (6.4.10)
Collecting qtconsole (from jupyter)
  Downloading http://mirrors.aliyun.com/pypi/packages/18/3b/7bdb99256d1406ae68904355222c6dc010640e6be158830f48b7097b
sole-5.4.3-py3-none-any.whl (121 kB)
Collecting jupyter-console (from jupyter)
  Downloading http://mirrors.aliyun.com/pypi/packages/ca/77/71d78d58f15c22db15328a476426f7ac4a60d3a5a7ba3b9627ee2f7903
er_console-6.5.3-py3-none-any.whl (24 kB)
Requirement already satisfied: nbconvert in /usr/local/lib/python3.8/dist-packages (from jupyter) (7.2.9)
Requirement already satisfied: ipykernel in /usr/local/lib/python3.8/dist-packages (from jupyter) (6.21.2)
Requirement already satisfied: ipywidgets in /usr/local/lib/python3.8/dist-packages (from jupyter) (8.0.6)
Requirement already satisfied: comm>=0.1.1 in /usr/local/lib/python3.8/dist-packages (from ipykernel>jupyter) (0.1.2)
Requirement already satisfied: debugpy>=1.6.5 in /usr/local/lib/python3.8/dist-packages (from ipykernel>jupyter) (1.6)
```

#### 5. 安装ipywidgets。

```
python3 -m pip install ipywidgets
```

```
root@48760329c039:/workspace# python3 -m pip install ipywidgets -l http://mirrors.aliyun.com/pypi/simple --trusted-hosts aliyun.com
Looking in indexes: http://mirrors.aliyun.com/pypi/simple, https://pypi.ngc.nvidia.com
Requirement already satisfied: ipywidgets in /usr/local/lib/python3.8/dist-packages (8.0.6)
Requirement already satisfied: ipykernel>=5.1 in /usr/local/lib/python3.8/dist-packages (from ipywidgets) (6.21.2)
Requirement already satisfied: ipython>=6.1.0 in /usr/local/lib/python3.8/dist-packages (from ipywidgets) (8.11.0)
Requirement already satisfied: traitlets>=4.3.1 in /usr/local/lib/python3.8/dist-packages (from ipywidgets) (5.9.0)
Requirement already satisfied: widgetsnbextension>=4.0.7 in /usr/local/lib/python3.8/dist-packages (from ipywidgets) (4.12.0)
Requirement already satisfied: jupyterlab-widgets>=3.0.7 in /usr/local/lib/python3.8/dist-packages (from ipywidgets) (3.0.7)
Requirement already satisfied: comm>=0.1.1 in /usr/local/lib/python3.8/dist-packages (from ipykernel>=4.5.1->ipywidget 2)
Requirement already satisfied: debugpy>=1.6.5 in /usr/local/lib/python3.8/dist-packages (from ipykernel>=4.5.1->ipywid 6.6)
Requirement already satisfied: jupyter-client>=6.1.12 in /usr/local/lib/python3.8/dist-packages (from ipykernel>=4.5.1 gets) (8.0.3)
Requirement already satisfied: jupyter-core>=5.0.*,>=4.12 in /usr/local/lib/python3.8/dist-packages (from ipykernel>=4 ywidgets) (5.2.0)
```

## 6. 打开Notebook扩展。

```
jupyter nbextension enable --py widgetsnbextension
```

```
root@48760329c039:/workspace# jupyter nbextension enable --py widgetsnbextension
Enabling notebook extension jupyter-js-widgets/extension ...
  - Validating: OK
root@48760329c039:/workspace#
```

## 7. 安装jupyterlab。

```
python3 -m pip install jupyterlab
```

```
root@48760329c039:/workspace# python3 -m pip install jupyterlab -l http://mirrors.aliyun.com/pypi/simple --trusted-hosts aliyun.com
Looking in indexes: http://mirrors.aliyun.com/pypi/simple, https://pypi.ngc.nvidia.com
Requirement already satisfied: jupyterlab in /usr/local/lib/python3.8/dist-packages (2.3.2)
Requirement already satisfied: notebook>=4.3.1 in /usr/local/lib/python3.8/dist-packages (from jupyterlab) (6.4.10)
Requirement already satisfied: tornado>=6.0.0,<6.0.2 in /usr/local/lib/python3.8/dist-packages (from jupyterl )
Requirement already satisfied: jupyterlab-server<2.0,>1.1.5 in /usr/local/lib/python3.8/dist-packages (from jupyterla 0)
Requirement already satisfied: jinja2>=2.10 in /usr/local/lib/python3.8/dist-packages (from jupyterlab) (3.1.2)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.8/dist-packages (from jinja2>=2.10->jupyterla 2)
Requirement already satisfied: requests in /usr/local/lib/python3.8/dist-packages (from jupyterlab-server<2.0,>1.1.5-lab) (2.28.2)
Requirement already satisfied: json5 in /usr/local/lib/python3.8/dist-packages (from jupyterlab-server<2.0,>1.1.5->ju ) (0.9.11)
Requirement already satisfied: jsonschema>=3.0.1 in /usr/local/lib/python3.8/dist-packages (from jupyterlab-server<2.0 >jupyterlab) (4.17.3)
Requirement already satisfied: pyzmq>=17 in /usr/local/lib/python3.8/dist-packages (from notebook>=4.3.1->jupyterlab)
Requirement already satisfied: argon2-cffi in /usr/local/lib/python3.8/dist-packages (from notebook>=4.3.1->jupyterlab 0)
```

## 8. 打开jupyterlab服务扩展。

```
jupyter serverextension enable --py jupyterlab
```

```
root@48760329c039:/workspace# jupyter serverextension enable --py jupyterlab
Enabling: jupyterlab
  - Writing config: /root/.jupyter
    - Validating ...
      jupyterlab 2.3.2 OK
root@48760329c039:/workspace#
```

## 说明

若该步骤出现“ModuleNotFoundError: No module named 'pysqllite2'”报错，请执行以下步骤。

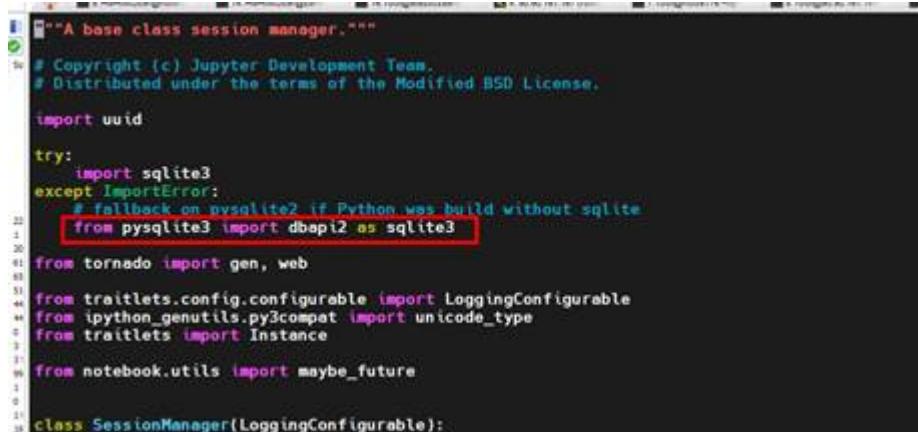
1. 执行如下命令。

```
apt-get install libsqlite3-dev  
pip3 install pysqllite3
```

2. 执行find / -name sessionmanager.py命令。

```
root@780c86625d52:/home/HwHiAiUser# find / -name sessionmanager.py  
/home/HwHiAiUser/.local/lib/python3.9/site-packages/notebook/services/sessions/sessionmanager.py  
/home/HwHiAiUser/.local/lib/python3.9/site-packages/jupyter_server/services/sessions/sessionmanager.py  
root@780c86625d52:/home/HwHiAiUser#
```

3. 将步骤2返回的两个文件中的from pysqllite2 import dbapi2 as sqlite3改为from pysqllite3 import dbapi2 as sqlite3。



```
"""A base class session manager."""  
# Copyright (c) Jupyter Development Team.  
# Distributed under the terms of the Modified BSD License.  
  
import uuid  
  
try:  
    import sqlite3  
except ImportError:  
    # fallback on pysqllite2 if Python was build without sqlite  
    from pysqllite3 import dbapi2 as sqlite3  
  
from tornado import gen, web  
  
from traitlets.config.configurable import LoggingConfigurable  
from ipython_genutils.py3compat import unicode_type  
from traitlets import Instance  
  
from notebook.utils import maybe_future  
  
class SessionManager(LoggingConfigurable):
```

9. jupyterlab安装完成后，需要修改默认工作目录，即Jupyterlab界面对应于容器中的哪个目录，此处设置的是/workspace。

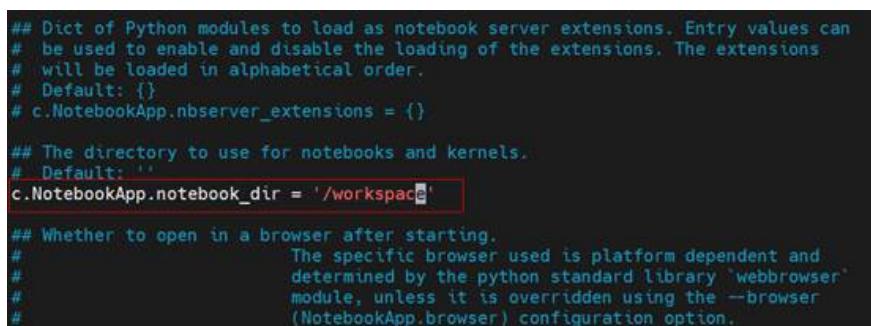
修改方法如下：

```
jupyter notebook --generate-config# 生成配置文件
```

```
root@a98ffca289a0:/workspace# jupyter notebook --generate-config  
Writing default config to: /root/.jupyter/jupyter_notebook_config.py  
root@a98ffca289a0:/workspace#
```

10. 根据提示路径打开jupyter\_notebook\_config.py文件，找到

```
#c.NotebookApp.notebook_dir = ; 去掉#，修改为c.NotebookApp.notebook_dir = '/workspace'。
```



```
## Dict of Python modules to load as notebook server extensions. Entry values can  
# be used to enable and disable the loading of the extensions. The extensions  
# will be loaded in alphabetical order.  
# Default: {}  
# c.NotebookApp.nbserver_extensions = {}  
  
## The directory to use for notebooks and kernels.  
# Default: ''  
c.NotebookApp.notebook_dir = '/workspace'  
  
## Whether to open in a browser after starting.  
# The specific browser used is platform dependent and  
# determined by the python standard library 'webbrowser'.  
#  
# module, unless it is overridden using the --browser  
# (NotebookApp.browser) configuration option.
```

## 步骤4 自动生成README.md。

用户进入jupyterlab界面后左侧会自动生成README.md文件，这种效果需要通过在容器中/etc/profile文件添加如下内容实现：

```
touch /workspace/README.md  
echo "如需使用本算法进行训练任务，需在要执行的Python文件中接收如下参数:"
```

data\_url: 数据集路径  
train\_out: 训练输出  
train\_log: 日志输出  
训练时这些参数名会以命令行传参的形式传入可执行的Python文件中，而参数值会由系统指定。例如：  
python run.py --key1=value1 ... --data\_url=/data/ --train\_out=/out/ --train\_log=/log/  
所以如果可执行的Python文件不接收这些参数值，可能会导致训练任务失败。" > /workspace/README.md

```
root@a98ffca289a0:/workspace# vi /etc/profile
root@a98ffca289a0:/workspace#
```

```
if [ -d /etc/profile.d ]; then
    for i in /etc/profile.d/*.sh; do
        if [ -r $i ]; then
            . $i
        fi
    done
    unset i
fi
touch /workspace/README.md
echo "如需使用本算法进行训练任务，需在要执行的python文件中接收如下参数：
data_url: 数据集路径
train_out: 训练输出
train_log: 日志输出
训练时这些参数名会以命令行传参的形式传入可执行的python文件中，而参数值会由系统指定。例如：
python run.py --key1=value1 ... --data_url=/data/ --train_out=/out/ --train_log=/log/
所以如果可执行的python文件不接收这些参数值，可能会导致训练任务失败。" > /workspace/README.md
```

## 步骤5 以上步骤执行完之后退出容器即可对镜像进行打包。

```
docker commit a98ffca289a0 jupyter_v2:v2
```

```
[root@node02 ~]# docker ps |grep jupyter_v2
a98ffca289a0      nvidia_chat              "/opt/nvidia/nvidia_..."   36 minutes ago   Up 36 minutes   6006/tcp,
8888/tcp  jupyter_v2
[root@node02 ~]# docker commit a98ffca289a0 jupyter_v2:v2
sha256:791a4c0d4378eff4d1fed95c809274d25a5dbb259e5c5907b071525e9240c558
```

## 步骤6 打包完成后需要用Dockerfile对该镜像进行一些处理用以完成README.md自动生成以及jupyterlab自动启动。

```
[root@node02 ~]# cd /home/image
[root@node02 image]# touch Dockerfile
[root@node02 image]# vi Dockerfile
```

新建一个文件，取名为Dockerfile，在Dockerfile中添加如下内容：

```
FROM jupyter_v2:v2 #FROM后面为上步骤打包的镜像名
# Expose Jupyter port & cmd
CMD /bin/bash -c "source /etc/profile && jupyter lab --ip=0.0.0.0 --port=8888 --no-browser --allow-root"

FROM jupyter_v2:v2
# Expose Jupyter port & cmd
CMD /bin/bash -c "source /etc/profile && jupyter lab --ip=0.0.0.0 --port=8888 --no-browser --allow-root"
```

在Dockerfile所在的目录执行：

```
docker build -t notebook01:v1 .
# notebook01:v1 名称可根据实际情况进行修改
```

```
[root@node02 image]# docker build -t notebook01:v1 .
Sending build context to Docker daemon 2.048kB
Step 1/2 : FROM jupyter_v2:v2
--> 791a4c0d4378
Step 2/2 : CMD /bin/bash -c "source /etc/profile && jupyter lab --ip=0.0.0.0 --port=8888 --no-browser --allow-root"
--> Running in 9ffadbed15b4
Removing intermediate container 9ffadbed15b4
--> b20aff5a669a
Successfully built b20aff5a669a
Successfully tagged notebook01:v1
[root@node02 image]#
```

镜像打包完成。

----结束

## 7.2. 昇腾镜像与非 root 用户镜像处理

- 若该Notebook需要调用NPU，则需要以昇腾镜像仓上的镜像为基础镜像，否则镜像需要手动安装CANN驱动，具体操作可参考昇腾官网。
- 若该Notebook无需调用NPU，则普通非root用户镜像（镜像执行用户必须是非root）等均需按照如下步骤处理。

本章节使用昇腾镜像infer-modelzoo来演示镜像打包的流程。

### 方案一：使用 Dockerfile 文件

Dockerfile示例：

```
FROM ascendhub.huawei.com/public-ascendhub/infer-modelzoo:23.0.RC2-mxvision
# 注意，该Dockerfile为昇腾样例，请按实际情况修改，具体步骤在方案二体现
USER root
RUN apt-get update && apt-get install libsqlite3-dev && usermod -g root HwHiAiUser && mkdir /workspace && \
echo"touch /workspace/README.md" >> /etc/profile && \
echo'echo "如需使用本算法进行训练任务，需在要执行的Python文件中接收如下参数：" >> /etc/profile && \
echo"data_url: 数据集路径" >> /etc/profile && \
echo"data_url: 数据集路径" >> /etc/profile && \
echo"train_log: 日志输出" >> /etc/profile && \
echo"训练时这些参数名会以命令行传参的形式传入可执行的Python文件中，而参数值会由系统指定。例如：" >> /etc/profile && \
echo"python run.py --key1=value1 .... --data_url=/data/ --train_out=/out/ --train_log=/log/" >> /etc/profile && \
echo'所以如果可执行的Python文件不接收这些参数值，可能会导致训练任务失败。" > /workspace/README.md' >> /etc/profile && \
chown HwHiAiUser:HwHiAiUser -R /workspace
USER HwHiAiUser
# Jupyterlab, 该镜像已存在Python3.9，具体镜像具体修改
RUN python3 -m pip install jupyter && python3 -m pip install ipywidgets && \
python3 -m pip install jupyter_contrib_nbextensions && python3 -m pip install notebook==6.1.0 && \
/home/HwHiAiUser/.local/bin/jupyter nbextension enable --py widgetsnbextension && \
python3 -m pip install jupyterlab && pip3 install pysqlite3 && \
sed -i "s|pysqlite2|pysqlite3|g" /home/HwHiAiUser/.local/lib/python3.9/site-packages/jupyter_server/services/sessions/sessionmanager.py && \
sed -i "s|pysqlite2|pysqlite3|g" /home/HwHiAiUser/.local/lib/python3.9/site-packages/notebook/services/sessions/sessionmanager.py && \
/home/HwHiAiUser/.local/bin/jupyter serverextension enable --py jupyterlab && \
/home/HwHiAiUser/.local/bin/jupyter notebook --generate-config && \
sed -i "#c.NotebookApp.notebook_dir = "|c.NotebookApp.notebook_dir = '/workspace'|g" /home/HwHiAiUser/.jupyter/jupyter_notebook_config.py
WORKDIR /workspace
CMD /bin/bash -c "source /etc/profile && jupyter lab --ip=0.0.0.0 --port=8888 --no-browser --allow-root"
```

### 方案二：docker commit 命令

**步骤1** 用此镜像起容器。

```
docker run -itd ascendhub.huawei.com/public-ascendhub/infer-modelzoo:23.0.RC2-mxvision /bin/bash
```

```
(base) [root@node176 xjh]# docker run -itd ascendhub.huawei.com/public-ascendhub/infer-modelzoo:23.0.RC2-mxvision /bin/bash
3813697a0b445b58c1e7d29ba69fa091ce4cc469117da98cca38b7a6f37718
(base) [root@node176 xjh]#
```

**步骤2** 查看容器。

```
docker ps
```

| CONTAINER ID | IMAGE  | COMMAND   | CREATED        | STATUS        | PO |
|--------------|--|-----------|----------------|---------------|----|
| 3813697a0b44 | ascendhub.huawei.com/public-ascendhub/infer-modelzoo:23.0.RC2-mxvision | /bin/bash | 20 seconds ago | Up 19 seconds |    |

**步骤3** 使用root用户，进入容器。

```
docker exec -u root -it 3813 bash
```

```
(base) [root@node176 xjh]# docker exec -u root -it 3813 bash
root@3813697a0b44:/home/HwHiAiUser#
```

**步骤4** 更新软件列表。

```
apt-get update
```

```
root@3813697a0b44:/home/HwHiAiUser# apt-get update
Get:1 https://repo.huaweicloud.com/ubuntu bionic InRelease [242 kB]
Get:2 https://repo.huaweicloud.com/ubuntu bionic-updates InRelease [88.7 kB]
Get:3 https://repo.huaweicloud.com/ubuntu bionic-backports InRelease [83.3 kB]
Get:4 https://repo.huaweicloud.com/ubuntu bionic-security InRelease [88.7 kB]
9% [Waiting for headers]
```

**步骤5** 将HwHiAiUser加入到root属组中

```
usermod -g root HwHiAiUser
```

```
root@3813697a0b44:/home/HwHiAiUser# usermod -g root HwHiAiUser
root@3813697a0b44:/home/HwHiAiUser#
```

**步骤6** 安装libsqllite3-dev。

```
apt-get install libsqllite3-dev
```

```
root@3813697a0b44:/home/HwHiAiUser# apt-get install libsqllite3-dev
Reading package lists... Done
Building dependency tree
Reading state information... Done
Suggested packages:
  sqlite3-doc
The following NEW packages will be installed:
  libsqllite3-dev
0 upgraded, 1 newly installed, 0 to remove and 56 not upgraded.
Need to get 633 kB of archives.
After this operation, 2136 kB of additional disk space will be used.
Get:1 https://repo.huaweicloud.com/ubuntu bionic-updates/main amd64 libsqllite3-dev amd64 3.22.0-1ubuntu0.7 [633 kB]
Fetched 633 kB in 3s (228 kB/s)
debconf: delaying package configuration, since apt-utils is not installed
Selecting previously unselected package libsqllite3-dev:amd64.
(Reading database ... 16533 files and directories currently installed.)
Preparing to unpack .../libsqllite3-dev_3.22.0-1ubuntu0.7_amd64.deb ...
Unpacking libsqllite3-dev:amd64 (3.22.0-1ubuntu0.7) ...
Setting up libsqllite3-dev:amd64 (3.22.0-1ubuntu0.7) ...
root@3813697a0b44:/home/HwHiAiUser#
```

**步骤7** 创建/workspace目录，并赋予HwHiAiUser用户权限。

```
mkdir /workspace
chown HwHiAiUser:HwHiAiUser -R /workspace
```

```
root@3813697a0b44:/home/HwHiAiUser# mkdir /workspace
root@3813697a0b44:/home/HwHiAiUser# chown HwHiAiUser:HwHiAiUser -R /workspace
```

**步骤8** 切换至HwHiAiUser用户。

```
su HwHiAiUser
```

```
root@3813697a0b44:/home/HwHiAiUser# su HwHiAiUser
HwHiAiUser@3813697a0b44:~$
```

**步骤9** 查看是否存在Python3。

```
HwHiAiUser@3813697a0b44:~$ python3
Python 3.9.2 (default, Aug 16 2023, 07:40:54)
[GCC 7.5.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

若无Python3，则需执行以下命令。

1. 安装Python。

```
apt-get install -y python3-dev
```

## 2. 安装pip。

```
apt-get install -y python3-pip
```

### 步骤10 安装jupyter。

```
python3 -m pip install jupyter
```

```
HwHiAiUser@3813697a0b44:~$ python3 -m pip install jupyter -i https://pypi.tuna.tsinghua.edu.cn/simple --trusted-host pypi.tuna.tsinghua.edu.cn
Defaulting to user installation because normal site-packages is not writeable
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting jupyter
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/83/df/9f5dd132200728e86190397e1e87cd76244e42d39ec5e88efd25b2abd7e/jupyter-1.0.0-py2.py3-none-any.whl (2.7 kB)
Collecting notebook (from jupyter)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/f3/2b/b984c57709b83c5cb818d21040db36719287f3d17db9b124c60cd483d94/noteboo
k-7.0.0-py3-none-any.whl (4.0 kB)
    1.8/4.0 MB 1.2 MB/s eta 0:00:02
```

### 步骤11 安装ipywidgets。

```
python3 -m pip install ipywidgets
```

```
HwHiAiUser@3813697a0b44:~$ python3 -m pip install ipywidgets -i https://pypi.tuna.tsinghua.edu.cn/simple --trusted-host pypi.tuna.tsinghua.edu.cn
Defaulting to user installation because normal site-packages is not writeable
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Requirement already satisfied: ipywidgets in ./local/lib/python3.9/site-packages (8.1.1)
Requirement already satisfied: comm>=0.1.3 in ./local/lib/python3.9/site-packages (from ipywidgets) (0.2.0)
Requirement already satisfied: ipython>=6.1.0 in ./local/lib/python3.9/site-packages (from ipywidgets) (8.18.1)
Requirement already satisfied: traitslet>=4.3.1 in ./local/lib/python3.9/site-packages (from ipywidgets) (5.14.0)
Requirement already satisfied: widgetsnbextension>=4.0.9 in ./local/lib/python3.9/site-packages (from ipywidgets) (4.0.9)
Requirement already satisfied: jupyterlab-widgets>=3.0.9 in ./local/lib/python3.9/site-packages (from ipywidgets) (3.0.9)
```

### 步骤12 因为是非root用户安装的jupyter，所以需要查看jupyter目录位置。

```
find / -name jupyter
```

```
HwHiAiUser@3813697a0b44:~$ find / -name jupyter
find: '/etc/ssl/private': Permission denied
find: '/proc/tty/driver': Permission denied
find: '/proc/1/mmap_files': Permission denied
find: '/proc/16/task/16/fd': Permission denied
find: '/proc/16/task/16/fdinfo': Permission denied
find: '/proc/16/task/16/ns': Permission denied
find: '/proc/16/fd': Permission denied
find: '/proc/16/mmap_files': Permission denied
find: '/proc/16/fdinfo': Permission denied
find: '/proc/16/ns': Permission denied
find: '/proc/259/task/259/fd': Permission denied
find: '/proc/259/task/259/fdinfo': Permission denied
find: '/proc/259/task/259/ns': Permission denied
find: '/proc/259/fd': Permission denied
find: '/proc/259/mmap_files': Permission denied
find: '/proc/259/fdinfo': Permission denied
find: '/proc/259/ns': Permission denied
find: '/var/lib/apt/lists/partial': Permission denied
find: '/var/cache/apt/archives/partial': Permission denied
find: '/var/cache/ldconfig': Permission denied
find: '/root': Permission denied
find: '/usr/local/Ascend/ascend-toolkit/6.3.RC2/mindstudio-toolkit/script': Permission denied
find: '/usr/local/Ascend/ascend-toolkit/6.3.RC2/toolkit/script': Permission denied
find: '/usr/local/Ascend/ascend-toolkit/6.3.RC2/tools/aoe/script': Permission denied
find: '/usr/local/Ascend/ascend-toolkit/6.3.RC2/tools/ncs/script': Permission denied
/home/HwHiAiUser/.local/bin/jupyter
/home/HwHiAiUser/.local/etc/jupyter
/home/HwHiAiUser/.local/share/jupyter
HwHiAiUser@3813697a0b44:~$
```

### 步骤13 安装jupyter\_contrib\_nbextensions。

```
python3 -m pip install jupyter_contrib_nbextensions
```

```
HwHiAiUser@3813697a0b44:~$ python3 -m pip install jupyter_contrib_nbextensions -i https://pypi.tuna.tsinghua.edu.cn/simple --trusted-host pypi.tuna.tsinghua.edu.cn
Defaulting to user installation because normal site-packages is not writeable
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting jupyter_contrib_nbextensions
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/50/91/78cc4362611dbde2b0c0d068204aef1b8899d0459c50d8ff9dcac8c069791/jupyter_
_contrib_nbextensions-0.7.0.tar.gz (23.5 kB)
    23.5/23.5 kB 1.3 MB/s eta 0:00:00
  Preparing metadata ... done
Collecting ipython_genutils (from jupyter_contrib_nbextensions)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/fa/bc/9bd3b5c2b4774d5f33b2d544f1460be9df7df2fe42f352135381c347c69a/ipython_
genutils-0.2.0-py2.py3-none-any.whl (26 kB)
Collecting jupyter_contrib_core>=0.3.1 (from jupyter_contrib_nbextensions)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/50/94/8d37e5b49emalc8bf284c46f9b0257c1f3319a4ab88acbd401da2cab25e55/jupyter_
contrib-core-0.4.2.tar.gz (17 kB)
```

### 步骤14 安装Notebook 6.1.0。

```
python3 -m pip install notebook==6.1.0
```

```
HwHiAiUser@3813697a0b44:~$ python3 -m pip install notebook==6.1.0 -i https://pypi.tuna.tsinghua.edu.cn/simple --trusted-host pypi.tsinghua.edu.cn
Defaulting to user installation because normal site-packages is not writeable
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting notebook==6.1.0
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/43/64/3fa6aab801f2e9898b1443966d8ea58ece094fdb:e64d2c0d928857df3472/notebook-6.1.0-py3-none-any.whl (9.4 MB)
    9.4/9.4 MB 1.2 MB/s eta 0:00:00
Requirement already satisfied: jinja2 in ./local/lib/python3.9/site-packages (from notebook==6.1.0) (3.1.2)
```

### 步骤15 打开Notebook扩展。

```
/home/HwHiAiUser/.local/bin/jupyter nbextension enable --py widgetsnbextension
```

```
HwHiAiUser@3813697a0b44:~$ /home/HwHiAiUser/.local/bin/jupyter nbextension enable --py widgetsnbextension
Enabling notebook extension jupyter-js-widgets/extension...
  - Validating: OK
HwHiAiUser@3813697a0b44:~$
```

### 步骤16 安装jupyterlab。

```
python3 -m pip install jupyterlab
```

```
HwHiAiUser@3813697a0b44:~$ python3 -m pip install jupyterlab
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: jupyterlab in ./local/lib/python3.9/site-packages (4.0.9)
Requirement already satisfied: async-tru>=1.0.0 in ./local/lib/python3.9/site-packages (from jupyterlab) (2.0.4)
Requirement already satisfied: importlib-metadata>=4.8.3 in /usr/local/python3.9.2/lib/python3.9/site-packages (from jupyterlab) (6.8.6)
```

### 步骤17 打开jupyterlab服务扩展。

```
/home/HwHiAiUser/.local/bin/jupyter serverextension enable --py jupyterlab
```

```
HwHiAiUser@3813697a0b44:~$ /home/HwHiAiUser/.local/bin/jupyter serverextension enable --py jupyterlab
Enabling: jupyterlab
  - Writing config: /home/HwHiAiUser/.jupyter
    - Validating...
      jupyterlab 4.0.9 OK
HwHiAiUser@3813697a0b44:~$
```

## 说明

若该步骤出现“ModuleNotFoundError: No module named 'pysqLite2'”报错，请执行以下步骤。

1. 执行如下命令。

```
apt-get install libsqlite3-dev
pip3 install pysqLite3
```

2. 执行命令find / -name sessionmanager.py。

```
root@780c86625d52:/home/HwHiAiUser# find / -name sessionmanager.py
/home/HwHiAiUser/.local/lib/python3.9/site-packages/notebook/services/sessions/sessionmanager.py
/home/HwHiAiUser/.local/lib/python3.9/site-packages/jupyter_server/services/sessions/sessionmanager.py
root@780c86625d52:/home/HwHiAiUser#
```

3. 将步骤2返回的两个文件中的from pysqLite2 import dbapi2 as sqLite3改为from pysqLite3 import dbapi2 as sqLite3。

```
"""A base class session manager."""
# Copyright (c) Jupyter Development Team.
# Distributed under the terms of the Modified BSD License.

import uuid

try:
    import sqLite3
except ImportError:
    # fallback on mysqLite2 if Python was build without sqLite
    from pysqLite3 import dbapi2 as sqLite3

from tornado import gen, web

from traitlets.config.configurable import LoggingConfigurable
from ipython_genutils.py3compat import unicode_type
from traitlets import Instance

from notebook.utils import maybe_future

class SessionManager(LoggingConfigurable):
```

### 步骤18 jupyterlab安装完成后，需要修改默认工作目录，即Jupyterlab界面对应于容器中的哪个目录，此处设置的是/workspace。

修改方法如下：

```
/home/HwHiAiUser/.local/bin/jupyter notebook --generate-config
```

```
HwHiAiUser@3813697a0b44:~$ /home/HwHiAiUser/.local/bin/jupyter notebook --generate-config
Writing default config to: /home/HwHiAiUser/.jupyter/jupyter_notebook_config.py
HwHiAiUser@3813697a0b44:~$
```

**步骤19** 根据提示路径打开jupyter\_notebook\_config.py文件，找到

```
#c.NotebookApp.notebook_dir = ; 去掉#，修改为c.NotebookApp.notebook_dir = '/workspace'。
```

```
2 # Will be loaded in alphabetical (cat. or def.) order.
3 # Default: {}
4 # c.NotebookApp.nbserver_extensions = {}
5 #
6 ## The directory to use for notebooks and kernels.
7 # Default: ''
8 c.NotebookApp.notebook_dir = '/workspace'
9 ## Whether to open in a browser after starting.
```

**步骤20** 自动生成README.md。

用户进入jupyterlab界面后左侧会自动生成README.md文件，这种效果需要通过在容器中/etc/profile文件添加如下内容实现：

```
touch /workspace/README.md
echo "如需使用本算法进行训练任务，需在要执行的Python文件中接收如下参数：
data_url: 数据集路径
train_out: 训练输出
train_log: 日志输出
训练时这些参数名会以命令行传参的形式传入可执行的Python文件中，而参数值会由系统指定。例如：
python run.py --key1=value1 .... --data_url=/data/ --train_out=/out/ --train_log=/log/
所以如果可执行的Python文件不接收这些参数值，可能会导致训练任务失败。" > /workspace/README.md
```

**步骤21** 以上步骤执行完之后退出容器即可对镜像进行打包。

```
docker commit 3813697a0b44 modelzoo-notebook-serving-x86:v1
```

```
(base) [root@node176 xjh]# docker commit 3813697a0b44 modelzoo-notebook-serving-x86:v1
```

**步骤22** 打包完成后需要用Dockerfile对该镜像进行一些处理用以完成README.md自动生成以及jupyterlab自动启动。

```
vi Dockerfile
```

```
(base) [root@node176 xjh]# vi Dockerfile
```

新建一个文件，取名为Dockerfile，在Dockerfile中添加如下内容：

```
FROM modelzoo-notebook-serving-x86:v1
USER HwHiAiUser
CMD /bin/bash -c "source /etc/profile && /home/HwHiAiUser/.local/bin/jupyter lab --ip=0.0.0.0 --port=8888 --no-browser --allow-root"
```

在Dockerfile所在的目录执行：

```
docker build -t notebook01:v1 .
# notebook01:v1 名称可根据实际情况进行修改
```

```
(base) [root@node176 xjh]# docker build -t notebook01:v1 .
[+] Building 0.0s (5/5) FINISHED
=> [internal] load build definition from Dockerfile
=> => transferring Dockerfile: 2B9B
=> [internal] load .dockerignore
=> => transferring context: 2B
=> [internal] load metadata for docker.io/library/modelzoo-notebook-serving-x86:v1
=> CACHED [!/] FROM docker.io/library/modelzoo-notebook-serving-x86:v1
=> => exporting layers
=> => Writing image sha256:514b2950ad3ddbe3a33ea83ccdd2905dc4f46ee423ab18c2b21b4d02ab661c855
=> => naming to docker.io/library/notebook01:v1
(base) [root@node176 xjh]#
```

镜像打包完成。

----结束

### 3.1.4.3 算法管理

算法管理分为“我的算法”和“内置算法”两个页面，用于管理自定义开发和AI Space预置的算法。

## 算法管理操作流程

图 3-66 算法管理操作流程



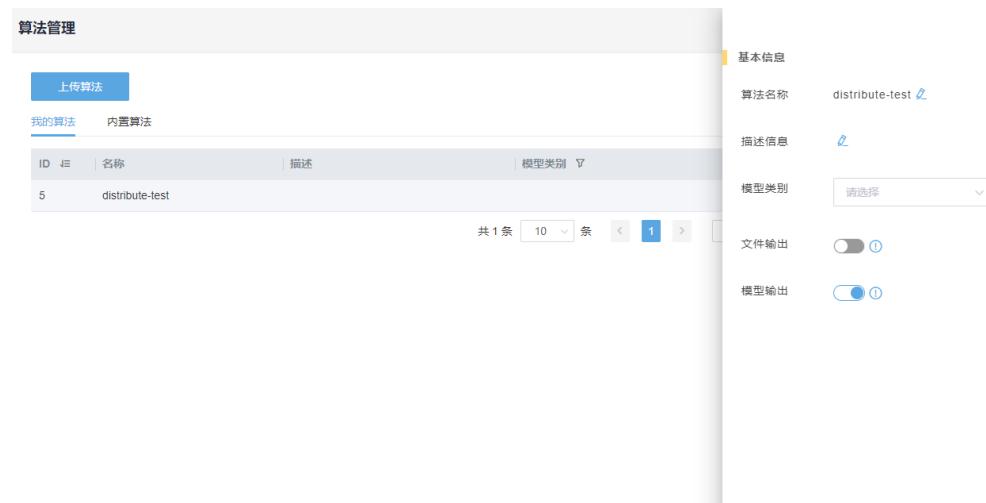
- 用户首先需要准备算法，算法的获取途径有四种：
  - 在“我的算法”页面中上传算法。
  - 使用“内置算法”页面中的内置算法。
  - Fork已有算法。
  - 从Notebook列表中保存算法。
- 我的算法列表中，可根据需要选择算法进行在线编辑，单击“在线编辑”后会进入JupyterLab页面，同时在Notebook列表中新增一条对应Notebook数据。编辑完成并保存算法后，可使用该算法创建训练任务，进入后续的训练环节。
- “我的算法”是用户创建的算法列表，可进行上传算法、在线编辑、创建训练任务、下载、Fork、删除等操作。
- “内置算法”是自带参数、命令、镜像等信息的算法列表，可快速创建训练任务，并可进行下载、Fork等操作，管理员权限支持上传、删除内置算法等操作。

## 查看算法详情

**步骤1** 选择“算法开发 > 算法管理”。

进入“算法管理”界面。

**步骤2** 单击列表中任意一条算法，在右侧弹窗中查看和编辑基本信息，包括编辑算法名称、描述信息、选择是否文件输出功能等。

**图 3-67 算法详情**

----结束

## 上传算法

**步骤1** 选择“算法开发 > 算法管理”。

进入“算法管理”界面。

**步骤2** 单击“上传算法”。

弹出“上传算法”提示框。

图 3-68 上传算法



步骤3 填下如下参数。

表 3-18 参数说明

| 参数   | 说明   |
|------|--|
| 名称   | 算法名称仅支持字母、数字、汉字、英文横杠和下划线，且算法名称不能重复。  |
| 描述   | 填写算法描述。  |
| 上传类别 | 勾选算法类型，包括： <ul style="list-style-type: none"><li>• 我的算法</li><li>• 内置算法</li></ul> |
| 模型类别 | 从下拉列表中选择模型类别。  |

| 参数    | 说明  |
|-------|---|
| 上传代码包 | <ul style="list-style-type: none"> <li>本地上传文件：请将算法文件打成一个zip包上传，单个文件不大于1GB。</li> <li>文件管理：单击“文件管理”，弹出文件管理弹窗，选择.zip包或者目录，单击“确定”。</li> </ul> <p><b>说明</b></p> <ul style="list-style-type: none"> <li>请确保代码中包含“train_model_out”参数用于接收训练的模型输出路径。</li> <li>如需断点续训或加载已有模型，请确保代码中包含“model_load_dir”参数用于接收训练模型路径。</li> <li>如需使用分布式训练，请确保代码中包含“num_nodes”参数和“node_ids”参数用于接收分布式相关参数。</li> <li>如需文件输出，请确保代码中包含“train_out”参数用于接收文件输出路径。</li> <li>如需使用GPU训练，请确保代码中包含“gpu_num_per_node”参数用于接收相关参数，后台将自动获取并填充。</li> <li>如需使用可视化训练，请指定可视化日志存放的路径为“/workspace/visualizedlog”。</li> </ul> |
| 镜像选择  | 算法来源选择内置算法时选择，选择镜像和镜像版本。  |
| 运行命令  | 算法来源选择内置算法时选择输入，输入运行命令，如python mnist.py。  |
| 文件输出  | 该算法若支持文件输出，请打开此开关。  |
| 模型输出  | 该算法若支持训练模型输出，请打开此开关。  |

**步骤4** 单击“确定”。

----结束

## 在线编辑算法

**步骤1** 选择“算法开发 > 算法管理”。

进入“算法管理”界面。

**步骤2** 选择需要在线编辑的算法，单击“在线编辑”。

弹出“设置规格”提示框。

图 3-69 设置规格



### 步骤3 设置计算类型和资源规格。

**步骤4** 单击“确定”。

后台会自动在Notebook列表中新建该算法对应的Notebook，并自动跳转到JupyterLab编辑页面。

## 说明

- 用户首次单击“在线编辑”时，后台自动创建Notebook并跳转JupyterLab需要时间，请耐心等待。
  - Notebook列表中新建的Notebook命名规则为：algorithm-{算法id}。

**步骤5** 在新打开的JupyterLab页面中，选择需要编辑的文件进行修改与调试。

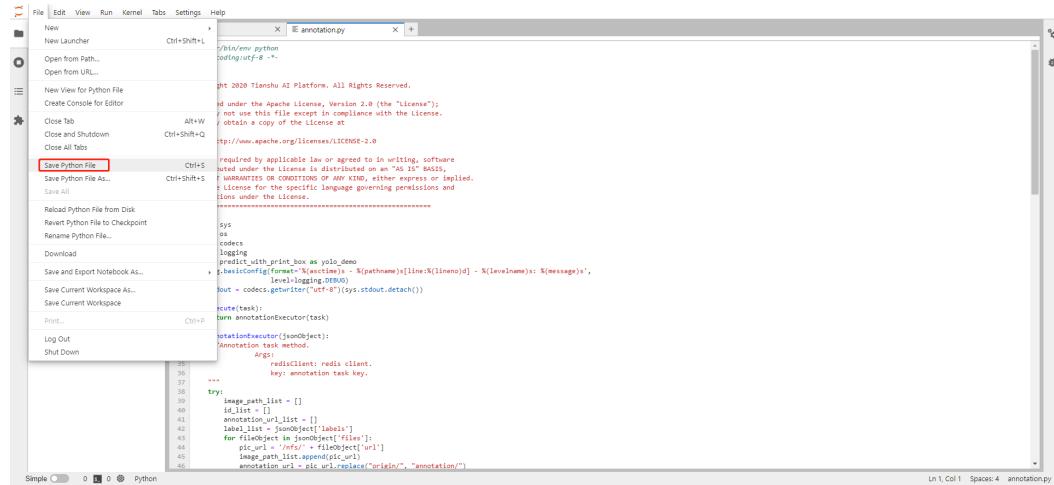
图 3-70 JupyterLab 页面

```
File Edit View Run Kernel Tabs Settings Help
Launcher x / annotation.py + 
Filter files by name
/
Name Last Modified
annotation... 7 days ago
annotation.py 7 days ago
inferenceny 7 days ago
predict_mlt... 7 days ago
README.md 2 months ago
yolo_npy 7 days ago

1 #!/usr/bin/env python
2 # -*- coding: utf-8 -*-
3 #
4 """
5 Copyright 2020 TianShu AI Platform. All Rights Reserved.
6 Licensed under the Apache License, Version 2.0 (the "License");
7 you may not use this file except in compliance with the License.
8 You may obtain a copy of the License at
9     http://www.apache.org/licenses/LICENSE-2.0
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License.
15 *****
16
17 import sys
18 import os
19 import redis
20 import logging
21 import predict
22 import json
23 import logging
24 import predict_with_redis as yolo_demo
25 logging.basicConfig(format='%(asctime)s : %(pathname)s(%(lineno)d) - %(levelname)s: %(message)s',
26                     level=logging.DEBUG)
27 sys.stdout = codecs.getwriter("utf-8")(sys.stdout.detach())
28
29 def execute(task):
30     return annotationExecutor(task)
31
32 def annotationExecutor(JsonObj):
33     """Annotation task method.
34     Args:
35         redisClient: redis client.
36         key: annotation task key.
37     """
38     try:
39         imgpathList = []
40         idList = []
41         annotationUrlList = []
42         labelList = []
43         labelsDict = {}
44         for fileObject in JsonObj['files']:
45             pic_url = '/nf/' + fileObject['url']
46             imgpathList.append(pic_url)
47             annotationUrlList.append(pic_url.replace('origin/', 'annotation/'))
48             idList.append(fileObject['id'])
49             labelList.append(fileObject['label'])
50             labelsDict[fileObject['label']] = fileObject['label']
51     except Exception as e:
52         print(e)
```

#### **步骤6** 修改完成后，保存算法文件。

图 3-71 保存修改



----结束

## 创建训练任务

### 说明

- 本节点说明的是在算法列表页面选择一个算法，使用该算法创建训练任务。
- 关于直接创建训练任务的说明，请参考[3.1.5.2 训练任务](#)章节。

**步骤1** 选择“算法开发 > 算法管理”。

进入“算法管理”界面。

**步骤2** 选择需要使用的算法，单击“创建训练任务”。

进入“创建任务”页面。

图 3-72 创建任务

The screenshot shows the 'Create Task' configuration page. It includes sections for 'Basic Information', 'Training Settings', and 'Parameter Mapping'.

**基本信息**

- 任务名称:
- 任务类型:
- 任务描述:

**训练设置**

- 算法类型:  我的算法  内置算法
- 算法名称:
- 选用图像:
- 模型类型:  我的模型  预训练模型
- 选择数据集:  训练数据集  验证数据集
- 运行命令:

**参数映射**

- 训练数据集:

### 步骤3 填写任务信息。

**表 3-19 参数说明**

| 参数   | 说明    |  |
|------|-------|--|
| 基本信息 | 任务名称  | 输入任务名称，支持字母、数字、汉字、英文横杠和下划线。  |
|      | 训练任务  | 训练任务类型。 <ul style="list-style-type: none"><li>● 单节点训练</li><li>● 多节点训练</li><li>● 大模型微调</li><li>● 大模型预训练</li></ul>   |
|      | 任务描述  | 输入任务描述。  |
| 训练设置 | 算法类型  | 选择算法的类型。 <ul style="list-style-type: none"><li>● 我的算法</li><li>● 内置算法</li></ul>   |
|      | 算法名称  | 从下拉列表中选择算法，内容会根据选用的算法类型变化。   |
|      | 选用镜像  | 从下拉列表中选择镜像和镜像版本。<br>支持TensorFlow、PyTorch等，镜像可在“镜像管理”界面查看。<br><b>说明</b><br>分布式训练场景下，用户需要在用于训练的镜像中预置以下软件： <ul style="list-style-type: none"><li>● 若镜像系统为Ubuntu，需要安装openSSH-server、jq、dnsutils。</li><li>● 若镜像系统为CentOS或RedHat，需要安装openSSH-server、jq、bind-utils。</li></ul> |
|      | 模型类型  | (可选) 选择模型类型，包括“我的模型”和“预训练模型”。  |
|      | 选用模型  | (选择模型类型后选择) 从下拉列表中选择模型和模型版本，模型可在“模型列表”界面查看。若选用则表示该模型可以作为本次训练的入参进行再次训练。   |
|      | 选择数据集 | 选择训练数据集、验证数据集的使用场景、数据集、数据集版本。<br>勾选验证数据集开启表示可以通过验证数据集校验训练出模型的推理精度。   |
|      | 运行命令  | 输入运行命令，如：python mnist.py。<br>运行命令目前只支持Python。  |
| 参数映射 | 训练数据集 | 如需传入训练数据集，请填写您的算法代码中用于接收训练数据集路径的参数。  |

| 参数    | 说明  |
|-------|---|
| 验证数据集 | 如需传入验证数据集，请填写您的算法代码中用于接收验证数据集路径的参数。   |
| 训练模型  | 如需断点续训或加载已有模型，请填写您的算法代码中用于接收训练模型路径的参数。  |
| 模型输出  | 如需输出模型，请填写您的算法代码中用于接收模型输出路径的参数。   |
| 资源规格  | <p>节点数</p> <ul style="list-style-type: none"> <li>任务类型为单节点训练、大模型微调的训练任务，节点数为1。</li> <li>任务类型为多节点训练、大模型预训练的训练任务，节点数大于等于2，表示本次训练将通过分布式模式运行，请确保代码中包含“num_nodes”参数和“node_ids”参数用于接收分布式相关参数。</li> </ul> <p>节点组</p> <p>从下拉列表中选择训练任务关联的节点组。</p> <p>训练加速框架</p> <p>若用户选择通过分布式模式进行训练，可从下拉列表中选择训练的加速框架。</p> <ul style="list-style-type: none"> <li>无</li> <li>DeepSpeed</li> <li>AscendSpeed / PTD模式</li> <li>Megatron-LM</li> <li>MindFormers</li> </ul> <p><b>说明</b></p> <ul style="list-style-type: none"> <li>DeepSpeed加速框架目前仅支持以下场景：在X86架构、Ubuntu 22.04操作系统上安装的AI Space上创建训练任务，集群中的Worker节点为H800满卡，且各Worker节点配置相同。</li> <li>仅调度器类型为Volcano时支持AscendSpeed训练加速框架。</li> <li>Megatron-LM框架适用于GPU整卡场景。</li> <li>MindFormers框架适用于NPU整卡场景。</li> </ul> <p>计算类型</p> <p>根据不同的算法选择需要的计算节点类型，包括：</p> <ul style="list-style-type: none"> <li>基础资源</li> <li>GPU / 整卡</li> <li>GPU / MIG</li> <li>NPU / 整卡</li> <li>NPU / vNPU</li> </ul> <p>基础资源</p> <p>从下拉框选择基础资源规格（资源规格可由超级管理员、组织管理员在控制台 &gt; 资源规格管理界面进行创建、修改）。</p> <p>选择完成后，将显示该基础资源规格参数信息，包括CPU个数、内存大小。</p> |

| 参数     | 说明  |
|--------|---|
|        | GPU/NPU规格<br>(当计算类型为GPU或NPU时需要选择)从下拉列表中选择资源规格(资源规格可由超级管理员、组织管理员在控制台>资源规格管理界面进行创建、修改)。若有多个节点，则显示每个节点的规格。<br>选择完成后，将显示该资源规格参数信息。  |
|        | 延迟启停<br>选择是否启动延迟启停，若用户启动延迟启停，可以进行如下设置： <ul style="list-style-type: none"><li>• 延迟启动：设置启动训练任务的延迟时间(以小时为单位)。</li><li>• 训练时长上限：设置训练任务运行最大时长(以小时为单位)。设置为0小时表示不限制训练时长。</li></ul> |
| 可视化    | 选用镜像<br>若用户需要创建可视化任务，选择可视化镜像和版本。  |
|        | 可视化资源规格<br>从下拉列表中选择可视化任务的资源规格。  |
| 运行命令预览 | -<br>运行命令的预览。   |

表 3-20 调度器配置

| 参数       | 说明   |
|----------|--|
| 调度器类型    | AI Space的调度器类型，超级管理员用户可在控制台>调度配置界面修改。<br>当调度器类型为Volcano时，填写下方调度参数。   |
| 所属队列     | 从下拉列表中选择训练任务所属队列。  |
| 优先级      | 选择训练任务的优先级。<br>当集群中运行多个job时，Volcano将以用户定义的优先级调度资源。   |
| 最少运行pod数 | 设置训练任务最少运行pod数，最少运行pod数需小于节点数。   |
| 最大重启次数   | 设置训练任务最大重启次数。  |
| 生命周期策略   | 选择Pod生命周期策略。支持“pod驱逐重启作业”，当pod被驱逐时，将重启该job。  |
| 节点亲和性    | 属性<br>输入节点属性，即节点标签，设置节点亲和性前需要为节点打上对应的标签。<br>单击  或  可添加或删除属性。 |

| 参数  | 说明  |
|-----|---|
| 类型  | 从下拉列表中选择节点亲和性类型，包括硬亲和性调度策略、软亲和性调度策略。 <ul style="list-style-type: none"><li>硬亲和性调度策略：调度器必须满足，多条规则间是一种“或”的关系，即只需要满足一条规则即会进行调度。</li><li>软亲和性调度策略：调度器会尽量满足，无论是满足其中一条或者是都不满足都会进行调度。</li></ul>                        |
| 操作符 | 从下拉列表中选择操作符。 <ul style="list-style-type: none"><li>In：标签的值在某个列表中</li><li>NotIn：标签的值不在某个列表中</li><li>Exists：某个标签存在</li><li>DoesNotExist：某个标签不存在</li><li>Gt：标签的值大于某个值（字符串比较）</li><li>Lt：标签的值小于某个值（字符串比较）</li></ul> |
| 取值  | 添加属性取值。<br>通过配置节点亲和性规则，调度器可以将Pod调度到具有特定标签的节点。<br>若用户选择DeepSpeed作为分布式训练框架，DeepSeep会为Pod添加反亲和性，每个节点只能调度一个Pod。   |

**步骤4** 单击“开始训练”。

----结束

## Fork 已有算法

用户在我的算法和内置算法列表中，可以通过Fork生成一份新算法。

**步骤1** 选择“算法开发 > 算法管理”。

进入“算法管理”页面。

**步骤2** 选择“我的算法”页签。

**步骤3** 在待操作的算法所在行，单击“更多 > Fork”。

弹出“fork生成算法”提示框。

**步骤4** 在弹出的窗口中填写信息。

### □□ 说明

Fork的算法名称不能与当前存在的算法名称重复。

**步骤5** 单击“确定”，算法创建完成，可在列表中查看。

----结束

## 下载算法

**步骤1** 选择“算法开发 > 算法管理”。

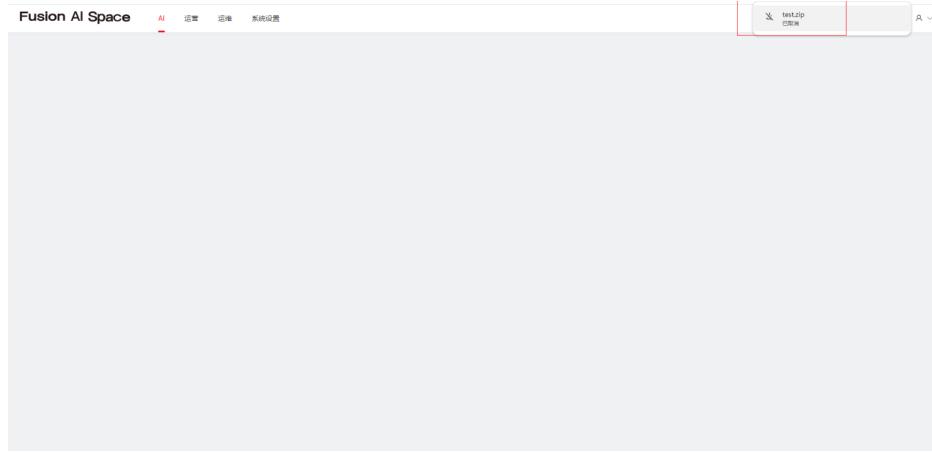
进入“算法管理”页面。

**步骤2** 选择“我的算法”页签。

**步骤3** 单击操作列的“更多 > 下载”，完成下载。

### 说明

下载文件时，如下图所示，可能会遭遇会话超时的问题，从而导致下载过程中断失败。为了解决这一问题，可以由超级管理员账户登录系统，在“系统设置 > 安全策略”界面中延长“会话超时时间”，设置完成后重新登录以下载文件。具体操作步骤请参见《AI Space 管理员指南》中的“安全策略”章节。



----结束

## 导出到文件管理

**步骤1** 选择“算法开发 > 算法管理”。

进入“算法管理”页面。

**步骤2** 选择“我的算法”页签。

**步骤3** 单击操作列的“更多 > 导出到文件管理”。

**步骤4** 弹出“文件管理”对话框。

**步骤5** 选择文件夹或新建文件夹，只能选择一个目标。

**步骤6** 单击“确定”，完成导出。

----结束

## 3.1.5 训练管理

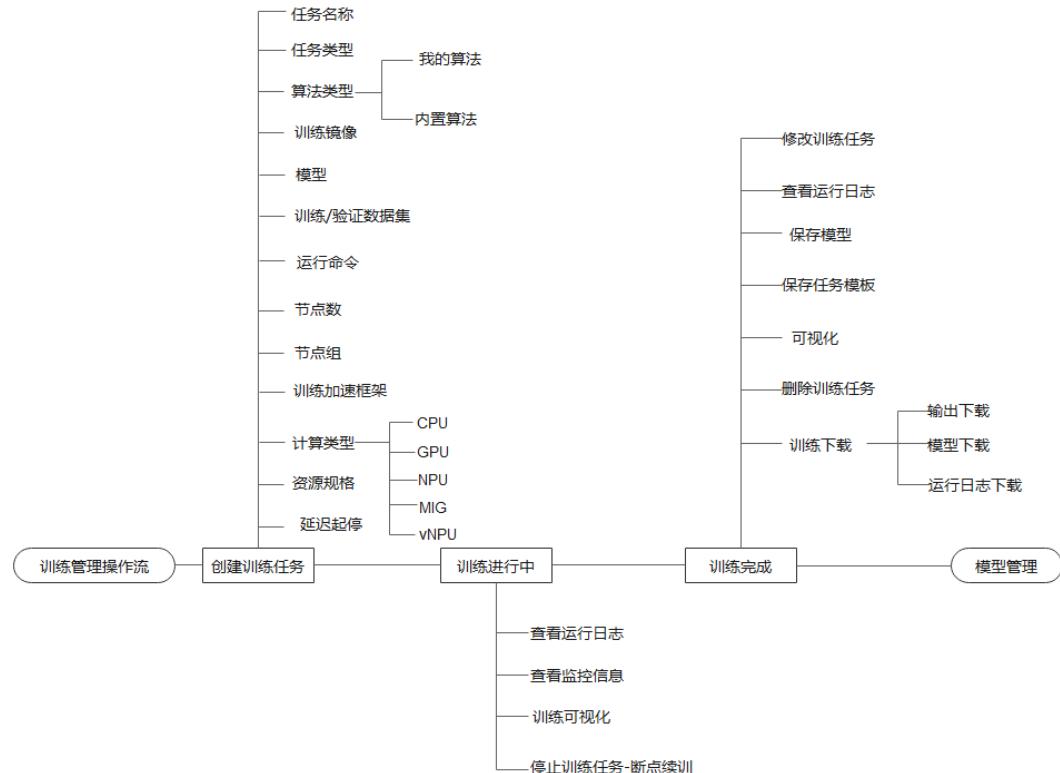
### 3.1.5.1 训练管理简介

AI Space为每一个训练任务都单独分配了一个虚拟容器去进行训练，各个训练任务之间相互隔离，互不干扰，提高了训练任务的可靠性。训练任务支持TensorFlow、

PyTorch等多种深度学习框架，使用内置算法或用户自定义算法进行云端训练。支持训练任务的多版本控制，用户可基于现有版本进行修改，动态调整算法超参数，从而得到一个满意的模型。

## 训练管理操作流程

图 3-73 训练管理流程图



## 昇腾镜像与非 root 用户镜像预处理

以昇腾镜像为例，Dockerfile如下：

```

FROM infer-modelzoo:23.0.RC2-mxvision
USER root
# 将普通用户加入root属组
RUN usermod -a -G root HwHiAiUser
# 切回普通用户
USER HwHiAiUser

```

### 3.1.5.2 训练任务

在“训练任务”界面，用户可以查看全部任务、运行中任务以及任务模板的相关信息，包括ID、名称、版本数等，并提供创建、删除、停止训练任务等操作。

- “全部任务”界面：可查看全部训练任务的ID、名称、任务类型、现有版本数目、训练时长、状态、创建时间等信息，并对任务进行创建、停止、删除等操作。
- “运行中任务”界面：仅显示状态为“运行中”的训练任务，可对任务进行停止等操作。

- “任务模板”界面：显示由状态为“运行完成”的训练任务保存的模板。

## 创建训练任务

### 说明

在创建训练任务之前，请确保已经准备好标注完成的数据集、可用的镜像和相应的算法。数据准备请参考[3.1.3.2.5 发布数据集](#)、[3.1.8.2 创建镜像](#)、[上传算法](#)章节。

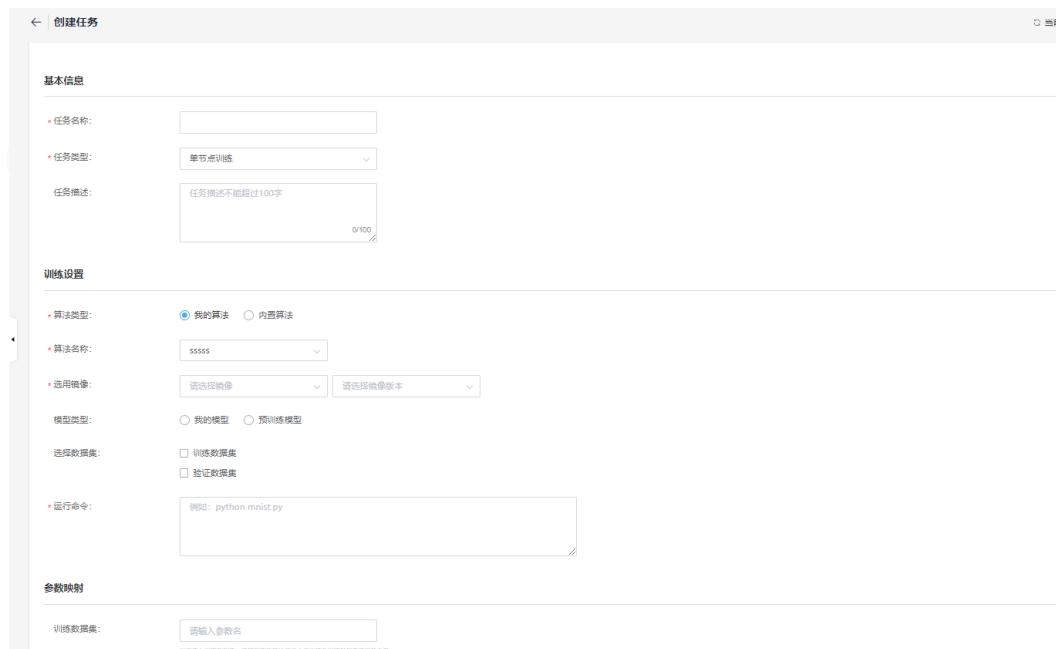
**步骤1** 选择“训练管理 > 训练任务”。

进入“训练任务”界面。

**步骤2** 单击“训练任务”下方的“创建”。

进入“创建任务”页面。

**图 3-74** 创建任务



**步骤3** 填写以下参数。

**表 3-21** 参数说明

| 参数   | 说明   |  |
|------|------|--|
| 基本信息 | 任务名称 | 输入任务名称，支持字母、数字、汉字、英文横杠和下划线。  |
|      | 训练任务 | 训练任务类型。 <ul style="list-style-type: none"><li>单节点训练</li><li>多节点训练</li><li>大模型微调</li><li>大模型预训练</li></ul> |

| 参数   |       | 说明  |
|------|-------|---|
|      | 任务描述  | 输入任务描述。   |
| 训练设置 | 算法类型  | <p>选择算法的类型。</p> <ul style="list-style-type: none"> <li>• 我的算法</li> <li>• 内置算法</li> </ul>  |
|      | 算法名称  | 从下拉列表中选择算法，内容会根据选用的算法类型变化。  |
|      | 选用镜像  | <p>从下拉列表中选择镜像和镜像版本。<br/>支持TensorFlow、PyTorch等，镜像可在“镜像管理”界面查看。</p> <p><b>说明</b><br/>分布式训练场景下，用户需要在用于训练的镜像中预置以下软件：</p> <ul style="list-style-type: none"> <li>• 若镜像系统为Ubuntu，需要安装openssh-server、jq、dnsutils。</li> <li>• 若镜像系统为CentOS或RedHat，需要安装opnssh-server、jq、bind-utils。</li> </ul> |
|      | 模型类型  | (可选) 选择模型类型，包括“我的模型”和“预训练模型”。   |
|      | 选用模型  | (选择模型类型后选择) 从下拉列表中选择模型和模型版本，模型可在“模型列表”界面查看。若选用则表示该模型可以作为本次训练的入参进行再次训练。  |
|      | 选择数据集 | <p>选择训练数据集、验证数据集的使用场景、数据集、数据集版本。</p> <p>勾选验证数据集开启表示可以通过验证数据集校验训练出模型的推理精度。</p>   |
|      | 运行命令  | <p>输入运行命令，如：python mnist.py。<br/>运行命令目前只支持Python。</p>   |
| 参数映射 | 训练数据集 | 如需传入训练数据集，请填写您的算法代码中用于接收训练数据集路径的参数。   |
|      | 验证数据集 | 如需传入验证数据集，请填写您的算法代码中用于接收验证数据集路径的参数。   |
|      | 训练模型  | 如需断点续训或加载已有模型，请填写您的算法代码中用于接收训练模型路径的参数。  |
|      | 模型输出  | 如需输出模型，请填写您的算法代码中用于接收模型输出路径的参数。   |

| 参数        | 说明  |
|-----------|---|
| 资源规格      | <p>节点数</p> <ul style="list-style-type: none"> <li>任务类型为单节点训练、大模型微调的训练任务，节点数为1。</li> <li>任务类型为多节点训练、大模型预训练的训练任务，节点数大于等于2，表示本次训练将通过分布式模式运行，请确保代码中包含“num_nodes”参数和“node_ids”参数用于接收分布式相关参数。</li> </ul>  |
| 节点组       | 从下拉列表中选择训练任务关联的节点组。   |
| 训练加速框架    | <p>若用户选择通过分布式模式进行训练，可从下拉列表中选择训练的加速框架。</p> <ul style="list-style-type: none"> <li>无</li> <li>DeepSpeed</li> <li>AscendSpeed / PTD模式</li> <li>Megatron-LM</li> <li>MindFormers</li> </ul> <p><b>说明</b></p> <ul style="list-style-type: none"> <li>DeepSpeed加速框架目前仅支持以下场景：在X86架构、Ubuntu 22.04操作系统上安装的AI Space上创建训练任务，集群中的Worker节点为H800满卡，且各Worker节点配置相同。</li> <li>仅调度器类型为Volcano时支持AscendSpeed训练加速框架。</li> <li>Megatron-LM框架适用于GPU整卡场景。</li> <li>MindFormers框架适用于NPU整卡场景。</li> </ul> |
| 计算类型      | 根据不同的算法选择需要的计算节点类型，包括： <ul style="list-style-type: none"> <li>基础资源</li> <li>GPU / 整卡</li> <li>GPU / MIG</li> <li>NPU / 整卡</li> <li>NPU / vNPU</li> </ul>  |
| 基础资源      | 从下拉框选择基础资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。选择完成后，将显示该基础资源规格参数信息，包括CPU个数、内存大小。  |
| GPU/NPU规格 | (当计算类型为GPU或NPU时需要选择) 从下拉列表中选择资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。若有多个节点，则显示每个节点的规格。选择完成后，将显示该资源规格参数信息。  |

| 参数     |         | 说明  |
|--------|---------|---|
|        | 延迟启停    | 选择是否启动延迟启停，若用户启动延迟启停，可以进行如下设置：<br><ul style="list-style-type: none"> <li>• 延迟启动：设置启动训练任务的延迟时间（以小时为单位）。</li> <li>• 训练时长上限：设置训练任务运行最大时长（以小时为单位）。设置为0小时表示不限制训练时长。</li> </ul> |
| 可视化    | 选用镜像    | 若用户需要创建可视化任务，选择可视化镜像和版本。  |
|        | 可视化资源规格 | 从下拉列表中选择可视化任务的资源规格。   |
| 运行命令预览 | -       | 运行命令的预览。  |

表 3-22 调度器配置

| 参数       |    | 说明   |
|----------|----|--|
| 调度器类型    |    | AI Space的调度器类型，超级管理员用户可在控制台 > 调度配置界面修改。<br>当调度器类型为Volcano时，填写下方调度参数。   |
| 所属队列     |    | 从下拉列表中选择训练任务所属队列。  |
| 优先级      |    | 选择训练任务的优先级。<br>当集群中运行多个job时，Volcano将以用户定义的优先级调度资源。   |
| 最少运行pod数 |    | 设置训练任务最少运行pod数，最少运行pod数需小于节点数。   |
| 最大重启次数   |    | 设置训练任务最大重启次数。  |
| 生命周期策略   |    | 选择Pod生命周期策略。支持“pod驱逐重启作业”，当pod被驱逐时，将重启该job。  |
| 节点亲和性    | 属性 | 输入节点属性，即节点标签，设置节点亲和性前需要为节点打上对应的标签。<br>单击  或  可添加或删除属性。 |
|          | 类型 | 从下拉列表中选择节点亲和性类型，包括硬亲和性调度策略、软亲和性调度策略。<br><ul style="list-style-type: none"> <li>• 硬亲和性调度策略：调度器必须满足，多条规则间是一种“或”的关系，即只需要满足一条规则即会进行调度。</li> <li>• 软亲和性调度策略：调度器会尽量满足，无论是满足其中一条或者是都不满足都会进行调度。</li> </ul>                           |

| 参数  |  | 说明   |
|-----|--|--|
| 操作符 |  | <p>从下拉列表中选择操作符。</p> <ul style="list-style-type: none"><li>In : 标签的值在某个列表中</li><li>NotIn : 标签的值不在某个列表中</li><li>Exists : 某个标签存在</li><li>DoesNotExist : 某个标签不存在</li><li>Gt : 标签的值大于某个值 (字符串比较)</li><li>Lt : 标签的值小于某个值 (字符串比较)</li></ul> |
| 取值  |  | <p>添加属性取值。</p> <p>通过配置节点亲和性规则，调度器可以将Pod调度到具有特定标签的节点。</p> <p>若用户选择DeepSpeed作为分布式训练框架，DeepSeep会为Pod添加反亲和性，每个节点只能调度一个Pod。</p>   |

**步骤4** 单击“开始训练”，将跳转至任务详情页。

单击“清空”将重置填写的参数。

----结束

## 任务版本详情

**步骤1** 选择“训练管理 > 训练任务”。

进入“训练任务”界面。

**步骤2** 选择需要查看任务版本详情的任务，单击任务名称。

进入任务详情页界面。

**图 3-75 训练任务**

The screenshot shows the 'Training Tasks' interface. At the top, there is a search bar with placeholder text '请输入任务名称或 ID' and a clear button. Below the search bar, there are three tabs: '全部任务' (All Tasks), '运行中任务' (Running Tasks), and '任务模板' (Task Templates). A large blue button labeled '新建' (Create New) is located at the top right. In the main area, there is a table with the following data:

| ID | 名称                               | 任务类型  | 现有版本数目 | 训练时长     | 状态   | 节点组 | 创建时间                | 操作                 |
|----|----------------------------------|-------|--------|----------|------|-----|---------------------|--------------------|
| 2  | train-6-20240517144422-v00<br>01 | 单节点训练 | 1      | 00:00:14 | 运行完成 |     | 2024-05-17 14:45:38 | <a href="#">删除</a> |
| 1  | sleep                            | 单节点训练 | 2      | 00:00:10 | 停止   |     | 2024-05-17 14:44:23 | <a href="#">删除</a> |

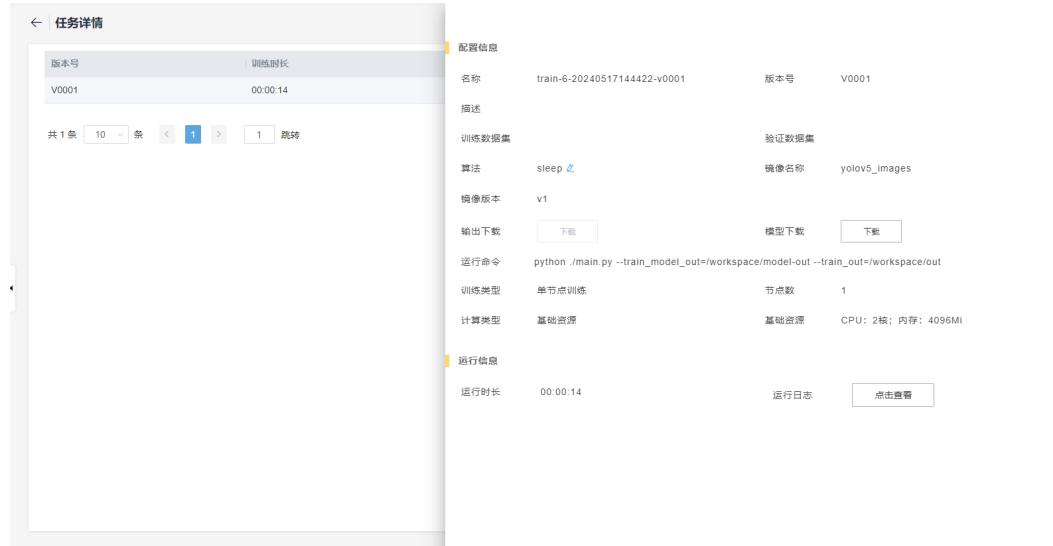
At the bottom of the table, there are pagination controls: '共 2 条' (2 items total), '10 条' (10 items per page), and buttons for '上一页' (Previous Page), '下一页' (Next Page), and '跳转' (Jump).

图 3-76 任务详情列表



**步骤3** 在任务详情界面选择一个版本，单击该记录条，在右侧弹出框中查看该版本的任务详情。

图 3-77 任务版本详情

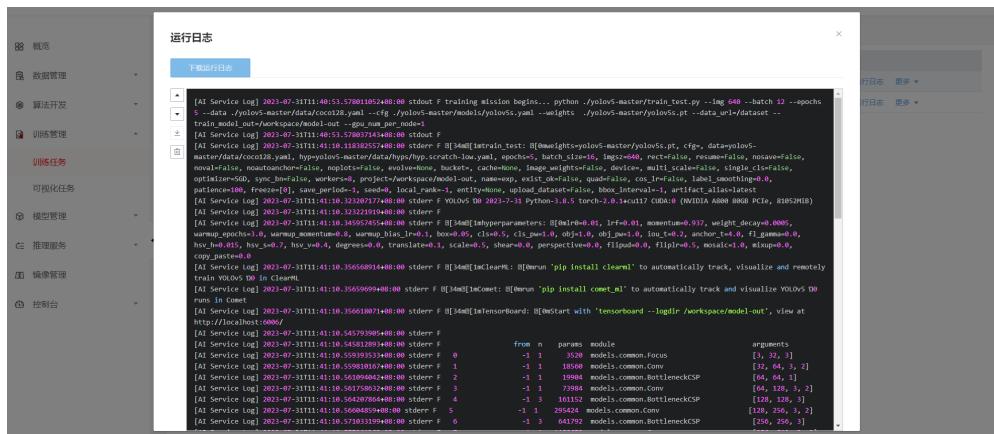


- 单击输出下载的“下载”，可下载该版本的训练日志。
- 单击模型下载的“下载”，可下载该版本训练生成的模型。
- 单击运行日志的“点击查看”，可在弹窗中查看训练日志。
- 在运行日志弹窗中，单击“下载运行日志”，可下载运行日志。

### 说明

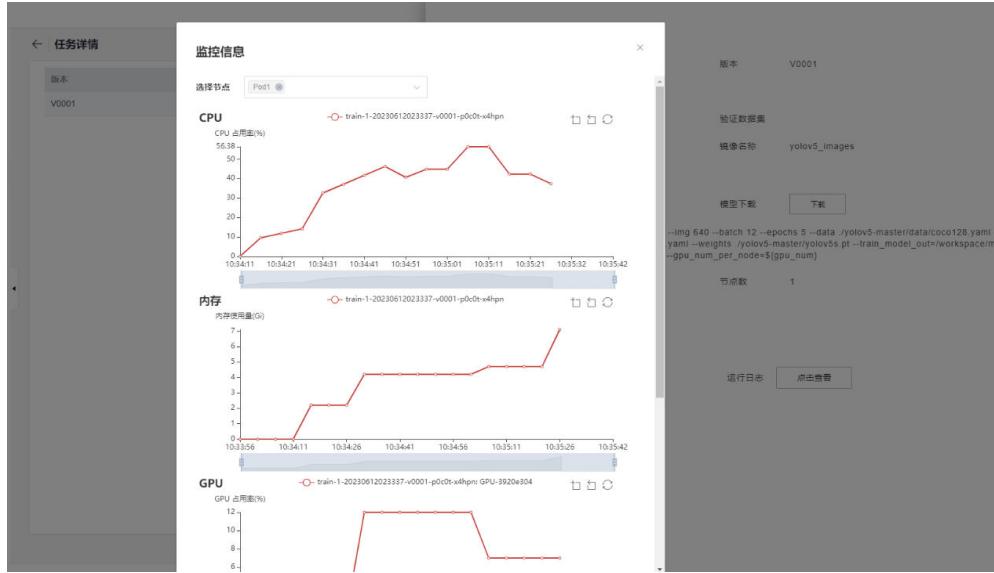
最多支持下载100MB大小的训练日志。

图 3-78 运行日志



- 运行中的任务，单击“查看监控信息”，可在弹窗中查看训练监控信息。

图 3-79 监控信息



#### 步骤4 任务详情界面支持操作。

- 单击操作栏的“修改”，可修改任务的各类参数并重新训练。
- 单击操作栏的“运行日志”，可查看运行日志。
- 单击操作栏的“更多 > 保存模型”，可保存模型。
- 单击操作栏的“更多 > 保存任务模板”，可将此版本的任务参数保存至任务模板列表页。
- 单击操作栏的“更多 > 删除”，该操作将删除此任务版本，且删除后无法恢复，请谨慎操作！

----结束

## 停止训练

对于状态为“运行中”任务支持停止操作。

#### 步骤1 选择“训练管理 > 训练任务”。

进入“训练任务”界面。

#### 步骤2 对于状态为“运行中”的任务，单击“停止”。

弹出提示框。

#### 步骤3 单击“确定”，稍后将自动停止训练，任务状态变为“停止”。

### 说明

单击训练任务列表上方的“一键停止”，可以停止所有正在运行中任务。

----结束

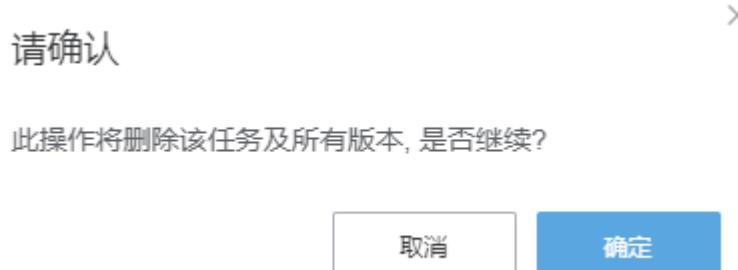
## 删除任务

**步骤1** 选择“训练管理 > 训练任务”。

进入“训练任务”界面。

**步骤2** 在待删除的任务所在行，单击“删除”。

弹出提示框。



**步骤3** 单击“确定”。

### 须知

“停止”、“运行完成”和“运行失败”任务支持删除操作，该操作将会删除该任务的所有版本且删除后无法恢复，请谨慎操作！

----结束

### 3.1.5.2.1 任务模板

“任务模板”保存训练运行时所需要的参数，方便用户快速使用。保存任务模板可以方便用户后续的参数修改扩展。

## 查看任务模板详情

**步骤1** 选择“训练管理 > 训练任务”。

进入“训练任务”界面。

**步骤2** 选择“任务模板”页签。

进入任务模板界面。

**步骤3** 在任务模板管理界面选择一条任务模板，单击记录条即可查看任务模板信息。

----结束

## 创建训练任务

**步骤1** 选择“训练管理 > 训练任务”。

进入“训练任务”界面。

**步骤2** 选择“任务模板”页签。

进入“任务模板”页面。

**步骤3** 在待创建训练任务的模板所在行，单击“创建训练任务”。

跳转至“创建任务”页面。

**图 3-80 用任务模板创建训练任务**

| ID | 任务模板名称                       | 任务类型  | 算法名称  | 数据集来源 | 节点类型 | 操作  |
|----|------------------------------|-------|-------|-------|------|---|
| 2  | train-5-20240515155052-v0001 | 单节点训练 | Ano   |       | CPU  | <a href="#">创建训练任务</a> <a href="#">编辑</a><br><a href="#">删除</a> |
| 1  | train-6-20240515154950-v0001 | 单节点训练 | sleep |       | CPU  | <a href="#">创建训练任务</a> <a href="#">编辑</a><br><a href="#">删除</a> |

**步骤4** 根据任务模板自动填写训练参数。

用户可以直接创建训练任务，亦可修改参数后创建。

#### 说明

关于创建训练任务的说明，请参考[创建训练任务](#)章节。

**步骤5** 单击“开始训练”。

----结束

## 编辑任务模板

**步骤1** 选择“训练管理 > 训练任务”。

进入“训练任务”界面。

**步骤2** 选择“任务模板”页签。

进入“任务模板”页面。

**步骤3** 在待编辑的任务模板所在行，单击“编辑”。

弹出“任务模板”提示框。

**步骤4** 对任务模板信息进行编辑修改。

**步骤5** 单击“确定”。

----结束

## 删除任务模板

**步骤1** 选择“训练管理 > 训练任务”。

进入“训练任务”界面。

**步骤2** 选择“任务模板”页签。

进入“任务模板”页面。

**步骤3** 在待删除的任务模板所在行，单击“删除”。

弹出提示框。



**步骤4** 单击“确定”。

----结束

### 3.1.5.3 可视化任务

#### 功能介绍

AI Space提供TensorBoard可视化功能，通过可视化界面展现模型结构、数据及参数，为深度学习模型训练及调优提供直观的参考，用户可以根据自己的需要创建多个可视化任务。

在“可视化任务”界面，用户可以查看可视化任务相关信息，包括名称、资源规格、状态、节点组、创建时间等操作，并提供启动、停止可视化任务的操作。

#### 创建可视化任务

**步骤1** 创建镜像。

- 准备好与算法版本匹配的TensorBoard镜像。

##### 说明

用户可在<https://hub.docker.com/search?q=tensorboard>自行下载TensorBoard镜像。

- 单击“镜像管理”，进入“镜像管理”页面。
- 参照**3.1.8.2 创建镜像**上传TensorBoard镜像，镜像用途选择可视化镜像。
- 单击“确定”。

**步骤2** 参考[上传算法](#)步骤上传算法。

##### 说明

在算法中指定可视化日志存放的路径为'/workspace/visualizedlog'。

**步骤3** 创建训练任务。

- 依次单击“训练管理 > 训练任务”，单击创建，进入“创建训练任务”界面。  
或在**步骤2**中刚上传完成的算法所在行，单击“创建训练任务”，进入“创建训练任务”页面。
- 参照[创建训练任务](#)填写训练任务信息。
  - 算法名称和**步骤2**中创建的算法保持一致。
  - 执行命令需要与算法文件中的命令保持一致。

- 可视化镜像选择**步骤1**中上传的可视化镜像。
  - 根据实际情况设置可视化任务的资源规格。
  - 根据实际情况设置节点数、节点类型、节点规格、节点组。
3. 设置完成后，单击“开始训练”。

**步骤4** 训练任务提交后，在可视化任务界面会同步创建一个名称相同、版本相同的可视化任务。

#### 📖 说明

最多支持10个版本的可视化任务。

#### ----结束

## 启动可视化

**步骤1** 依次单击“训练管理 > 可视化任务”。

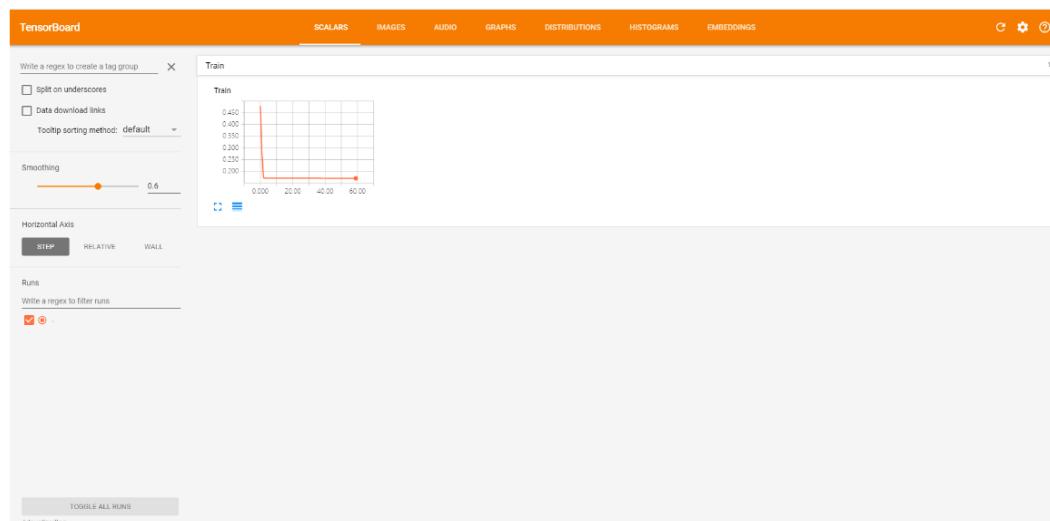
进入可视化界面。

**步骤2** 在需要启动的任务所在行，单击操作列的“启动”。

**步骤3** 启动可视化任务后，单击“打开”。

进入TensorBoard可视化界面查看训练任务的可视化结果。

图 3-81 可视化界面



**步骤4** 单击操作列的“停止”可以停止可视化。

#### ----结束

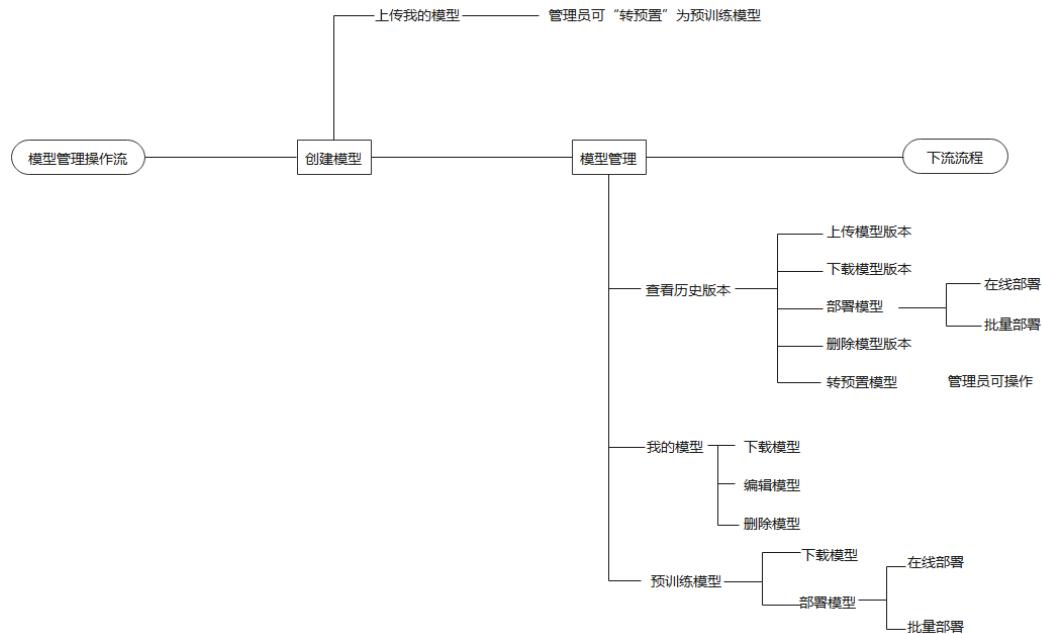
## 3.1.6 模型管理

### 3.1.6.1 模型管理简介

模型管理可以导入训练生成的模型，并对模型的版本迭代进行统一管理。用户可在此进行模型的管理，支持的操作包括创建模型、查询模型、历史版本、下载、编辑、删除等。

## 模型管理操作流程

图 3-82 模型管理流程图



### 3.1.6.2 模型列表

#### 3.1.6.2.1 我的模型

##### 上传模型

**步骤1** 选择“模型管理 > 模型列表”。

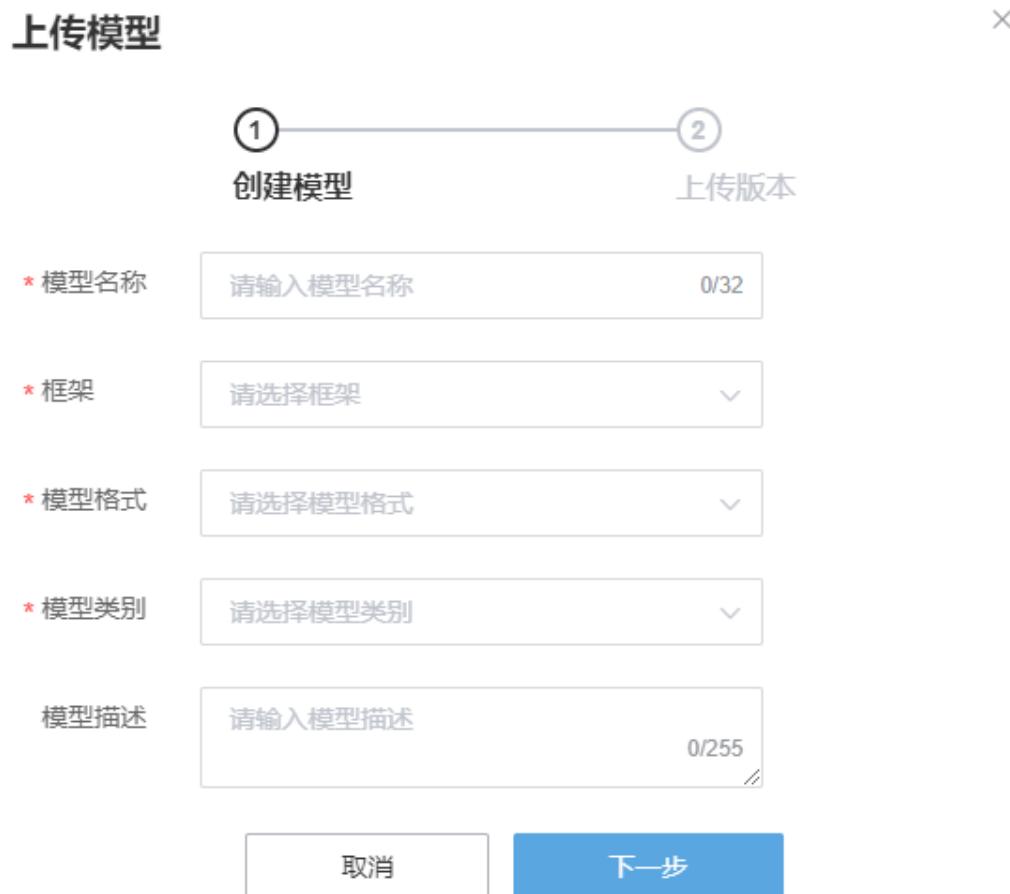
进入“模型列表”界面。

**步骤2** 选择“我的模型”页签。

**步骤3** 单击“上传模型”。

弹出“上传模型”提示框。

图 3-83 上传我的模型



**步骤4** 填写如下参数。

表 3-23 参数说明

| 参数   | 说明  |
|------|---|
| 模型名称 | 输入模型名称，支持字母、数字、汉字、英文横杠和下划线。   |
| 框架   | 从下拉列表中选择框架，包括TensorFlow、Pytorch、Keras、Caffe、Blade、Mxnet、Paddlepaddle、Mindspore。   |
| 模型格式 | 从下拉列表中选择模型格式，包括SaveModel、FrozenPb、KerasH5、CaffePrototxt、ONNX、BladeModel、PMML、Pytorch PTH、pb、ckpt、pkt、pt、HDF5、caffemodel、params、json、Directory等多种格式。 |
| 模型类别 | 从下拉列表中选择模型类别，包括图像分类、目标检测、语义分割、目标跟踪、文本分类、中文分词、命名实体识别、音频分类、语音识别、大模型和自定义模型。  |
| 模型描述 | 输入模型描述。   |

**步骤5** 单击“下一步”，当前页提示“模型新建成功”，进入上传版本页面。

图 3-84 上传版本

**步骤6** 文件上传。

- 选择“本地上传文件”，单击“上传文件”上传打包成.zip包的模型。
- 选择“文件管理”，单击“文件管理”，弹出文件管理界面，选择.zip格式压缩包或目录，单击“确定”。

**步骤7** 单击“确定上传”。**说明**

若模型暂时不上传版本信息，可单击“下次再传”按钮，返回至模型列表页，之后用户可在历史版本中上传模型。

**----结束**

## 模型版本管理

**步骤1** 选择“模型管理 > 模型列表”。

进入“模型列表”界面。

**步骤2** 选择“我的模型”页签。**步骤3** 在待操作模型所在行，单击“历史版本”。

进入“模型版本管理”页面。

图 3-85 模型列表

| 模型版本管理 |       |      |                               |                     |  | 当前资源占用量 —— CPU: 0核 / 0核 内存: 0Gi / 0Gi |
|--------|-------|------|-------------------------------|---------------------|--|---------------------------------------|
| 上传模型版本 |       |      |                               |                     |  |                                       |
| 序号     | 版本    | 模型来源 | 模型地址                          | 创建时间                | 操作   |                                       |
| 1      | V0001 | 用户上传 | /model/5/2024041616295009644t | 2024-04-16 16:30:03 | <a href="#">下载</a> <a href="#">在线服务</a> <a href="#">更多</a> |                                       |
| 共 1 条  | 10 条  | <    | 1                             | >                   | 筛选   |                                       |

**步骤4** 单击“上传模型版本”。

弹出“创建模型版本管理”提示框。

**图 3-86 上传模型版本**



**步骤5** 上传文件。

- 选择“本地上传文件”，单击“上传文件”上传打包成zip包的模型。
- 选择“文件管理”，单击“文件管理”，弹出文件管理界面，选择.zip格式压缩包或目录，单击“确定”。

**步骤6** 上传完成后，单击“确定”。

模型版本上传成功后，可对模型版本进行下载、部署、转预训练、删除等操作。

----结束

## 下载历史版本

**步骤1** 选择“模型管理 > 模型列表”。

进入“模型列表”界面。

**步骤2** 选择“我的模型”页签。

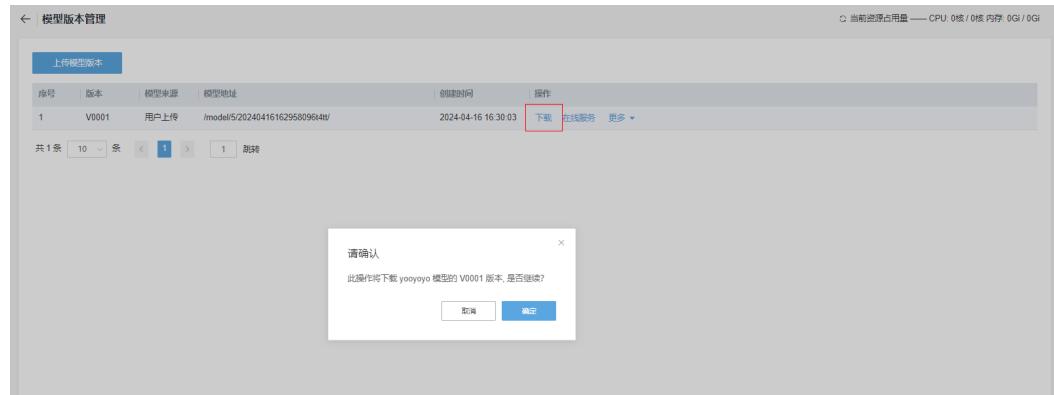
**步骤3** 在待下载历史版本的模型所在行，单击“历史版本”。

进入“模型版本管理”页面。

**步骤4** 选择需要下载的版本，单击“下载”。

弹出提示框。

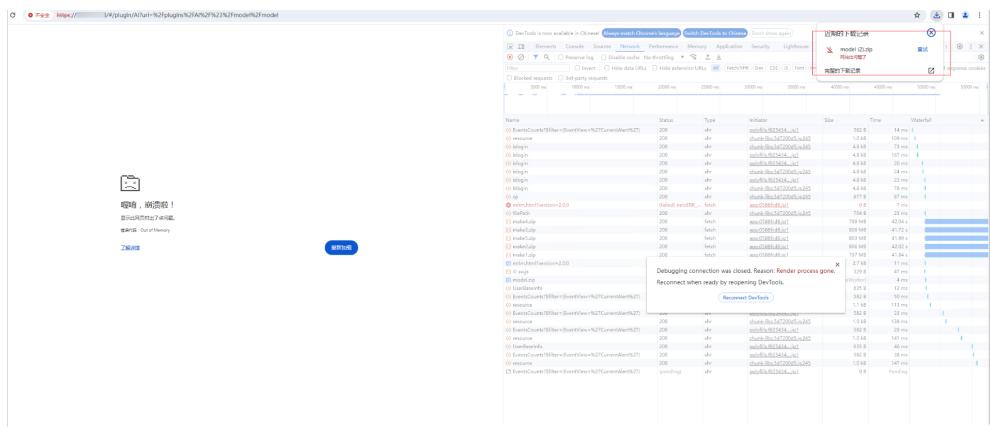
图 3-87 下载模型历史版本



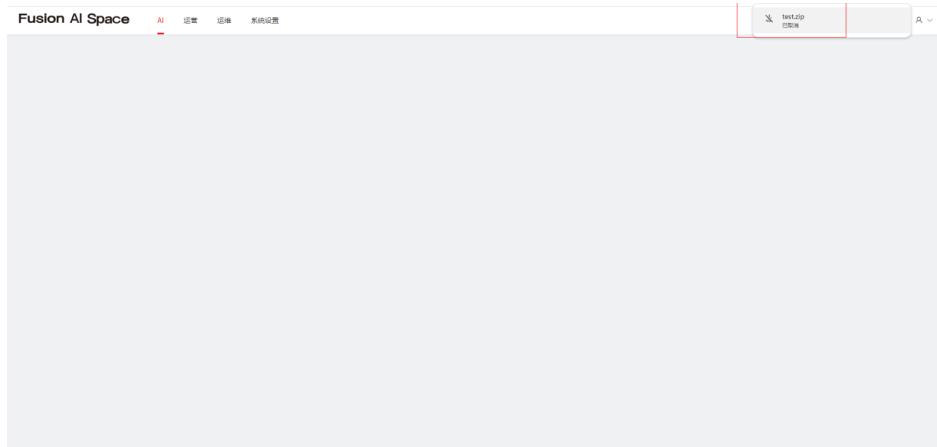
步骤5 单击“确定”。

#### 说明

- 若模型名称中包含中文字符，则下载模型至本地打开可能存在报错。
- 在下载过程中，如果如下图所示，出现界面崩溃情况，这可能与用户的个人设备相关，可以由超级管理员登录后台进行下载。若需要进一步的帮助，请参照**7 如何获取帮助**章节，联系技术支持团队。



- 下载文件时，如下图所示，可能会遭遇会话超时的问题，从而导致下载过程中断失败。为了解决这一问题，可以由超级管理员账户登录系统，在“系统设置 > 安全策略”界面中延长“会话超时时间”，设置完成后重新登录以下载文件。具体操作步骤请参见《AI Space 管理员指南》中的“安全策略”章节。



----结束

## 下载当前版本

- 步骤1** 选择“模型管理 > 模型列表”。
- 进入“模型列表”界面。
- 步骤2** 选择“我的模型”页签。
- 步骤3** 在待下载当前版本的模型所在行，单击“下载”。
- 弹出提示框。

图 3-88 下载模型当前版本



### 说明

如无模型版本信息，则下载按钮不可用。

- 步骤4** 单击“确定”。若模型名称中包含中文字符，则下载模型至本地打开可能存在报错。  
----结束

## 编辑模型

- 步骤1** 选择“模型管理 > 模型列表”。
- 进入“模型列表”界面。
- 步骤2** 选择“我的模型”页签。
- 步骤3** 在待编辑的模型所在行，单击“更多 > 编辑”，  
弹出“修改模型”提示框。

### 说明

未上传模型版本的模型，请单击操作列的“编辑”，对模型进行修改。

图 3-89 修改模型



**步骤4** 对模型信息进行编辑修改。

**步骤5** 单击“确定”。

----结束

## 删除模型

**步骤1** 选择“模型管理 > 模型列表”。

进入“模型列表”界面。

**步骤2** 选择“我的模型”页签。

**步骤3** 在待删除的模型所在行，单击“更多 > 删除”。

弹出提示框。

### 说明

未上传模型版本的模型，请单击操作列的“删除”，对模型进行删除。

图 3-90 删除模型

| 模型列表 |          |         |             |      |      |       |       |                     |            | 当前资源占用量 —— CPU: 4核 / 2048核 GPU: 0卡 / 4卡 NPU: 0卡 / 1卡 vNPU: 0卡 / 1卡 内存: 8.0Gi / 4196.0Gi |    |    |
|------|----------|---------|-------------|------|------|-------|-------|---------------------|------------|---|----|----|
| 我的模型 |          |         |             |      |      |       |       |                     |            | 搜索输入模型名称  |    |    |
| 序号   | 模型名称     | 框架      | 模型格式        | 模型类别 | 模型描述 | 版本    | 创建者ID | 更新时间                | 操作         | 历史版本  | 下载 | 更多 |
| 1    | true1    | pytorch | Pytorch PTH | 图像分类 |      | V0001 | 5     | 2024-06-18 15:35:18 | 历史版本 下载 更多 |   |    |    |
| 2    | serv     | pytorch | Pytorch PTH | 图像分类 |      | V0001 | 5     | 2024-06-18 14:41:23 | 历史版本       | 删除  |    |    |
| 3    | v_123d-  | pytorch | Pytorch PTH | 图像分类 |      | V0001 | 5     | 2024-06-17 10:56:29 | 历史版本       | 删除  |    |    |
| 4    | classify | pytorch | Pytorch PTH | 图像分类 |      | V0001 | 5     | 2024-06-15 14:07:35 | 历史版本       | 导出到文件管理   |    |    |
| 5    | model    | pytorch | Pytorch PTH | 大模型  |      | V0004 | 5     | 2024-06-15 11:34:27 | 历史版本       | 导出到文件管理   |    |    |

**步骤4** 单击“确定”。

#### 须知

所选模型信息以及该模型的所有版本将被删除且无法恢复，请谨慎操作。

----结束

### 导出到文件管理

**步骤1** 选择“模型管理 > 模型列表”。

进入“模型列表”界面。

**步骤2** 选择“我的模型”页签。

**步骤3** 在目标模型所在行，单击“历史版本”。

进入“模型版本管理”页面。

**步骤4** 单击操作列的“更多 > 导出到文件管理”。

**步骤5** 弹出“文件管理”对话框。

**步骤6** 选择文件夹或创建新文件夹，只能选择一个目标。

**步骤7** 单击“确定”，完成导出。

----结束

### 3.1.6.2.2 预训练模型

组织管理员可将“我的模型”通过“转预训练”功能转化为“预训练模型”。

### 转预训练模型

**步骤1** 选择“模型管理 > 模型列表”。

进入“模型列表”界面。

**步骤2** 选择“我的模型”页签。

**步骤3** 在待转预训练的模型所在行，单击“历史版本”。

进入“模型版本管理”页面。

**步骤4** 在待转预训练的模型版本所在行，单击“更多 > 转预训练”。

弹出提示框。

图 3-91 转为预训练模型



**步骤5** 单击“确定”，模型将展示在预训练模型列表。

----结束

## 下载预训练模型

**步骤1** 选择“模型管理 > 模型列表”。

进入“模型列表”界面。

**步骤2** 选择“预训练模型”页签。

**步骤3** 在待下载的模型所在行，单击“下载”。

弹出提示框。



**步骤4** 单击“确定”。若模型名称中包含中文字符，则下载模型至本地打开可能存在报错。

----结束

## 删除预训练模型

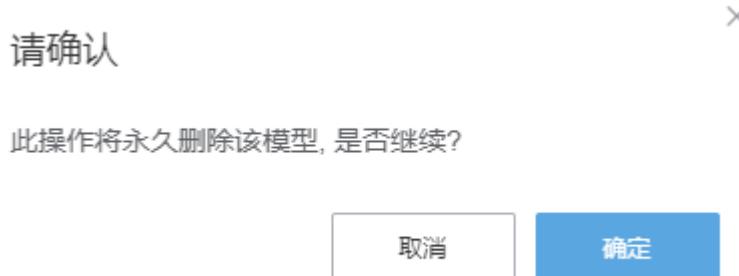
**步骤1** 选择“模型管理 > 模型列表”。

进入“模型列表”界面。

**步骤2** 选择“预训练模型”页签。

**步骤3** 在待删除的模型所在行，单击“更多 > 删除”。

弹出提示框。



**步骤4** 单击“确定”。

----结束

## 3.1.7 推理服务

### 3.1.7.1 推理服务简介

完成模型训练及模型格式转化后，模型管理中保存了指定格式模型。用户可在推理服务模块对这些模型进行部署或不选择模型直接部署。

#### 推理服务框架特点

推理服务支持平台框架和用户自定义框架。两种推理服务框架特点如表3-24所示。

**表 3-24 框架特点**

| 服务框架类型 | 特点  |
|--------|---|
| 平台框架   | <ul style="list-style-type: none"><li>支持使用自制镜像。</li><li>支持对TensorFlow，PyTorch两种深度学习框架训练的模型部署。</li><li>支持不选择模型部署。</li><li>支持HTTP、GRPC两种通信方式。</li><li>支持多节点部署。</li><li>支持灰度发布。</li><li>支持基本的图片预处理。</li><li>支持自定义推理脚本。</li><li>支持使用自制镜像。自制镜像需参考<a href="#">3.1.7.2 推理镜像预处理（平台框架）</a>章节安装平台框架所需相关依赖，并且根据用户业务，安装业务所需的依赖。</li></ul> |

| 服务框架类型 | 特点   |
|--------|--|
| 自定义框架  | <ul style="list-style-type: none"><li>支持用户部署自定义推理镜像。</li><li>支持对TensorFlow，PyTorch两种深度学习框架训练的模型部署。</li><li>支持不选择模型部署。</li><li>支持HTTP通信方式。</li><li>支持多节点部署。</li><li>支持灰度发布。</li></ul> |

## 推理服务方式

推理服务支持以下两种方式进行模型部署：

- 在线推理**：将模型部署为在线服务，通过API的方式为用户提供推理预测能力。
- 批量推理**：可以对批量数据进行推理，推理完成后服务结束。

## 推理服务支持的模型结构

表 3-25

| 服务类型 | 服务架构 | 支持的模型结构  |
|------|------|--|
| 在线推理 | 自定义  | 支持SaveModel、FrozenPb、KerasH5、CaffePrototxt、ONNX、BladeModel、PMML、Pytorch PTH、pb、ckpt、pkt、pt、h5 ( HDF5 ) 、caffemodel、params、json、Directory、pdparams、pdiparams、pdmodel多种格式模型。 |
|      | 平台   | <ul style="list-style-type: none"><li>Tensorflow框架的Savedmodel格式保存的模型</li><li>Pytorch支持内置模型.pth格式的模型</li><li>Keras框架的Savedmodel格式保存的模型</li></ul>                            |
| 批量推理 | 自定义  |  |
|      | 平台   |  |

## 推理服务操作流程

图 3-92 推理服务流程图



### 3.1.7.2 推理镜像预处理（平台框架）

若用户部署推理服务时使用平台框架，对应的镜像需要满足以下条件。

- Python软链接为镜像中的Python3。如通过命令映射Python软链接至Python3。  
`mv /usr/bin/python /usr/bin/python_bak  
ln -s /usr/local/python3.7.5/bin/python3 /usr/bin/python`
- Python需有推理框架的依赖，才能在AI Space运行推理服务，安装步骤请参见[安装Python推理框架依赖](#)。
- 若镜像启动的默认用户为非root用户，则需要将该用户加入root属组，操作步骤请参见[将用户加入root属组](#)。

### 安装 Python 推理框架依赖

**步骤1** 执行`vi serving-requirements.txt`命令。

**步骤2** 按*i*进入编辑，将以下内容复制至文件内。

```

click
charset-normalizer
fastapi
grpcio
h11
idna
pydantic
requests
six
starlette
typing_extensions
  
```

```
urllib3
uvicorn
wincertstore
python-multipart
google-auth
google-auth-oauthlib
google-pasta
protobuf
```

**步骤3** 按ESC键，输入“:wq”保存文件。

**步骤4** 执行如下命令安装依赖。

```
pip3 install -r serving-requirements.txt
```

----结束

## 将用户加入 root 属组

**步骤1** 切换到root用户。

```
su - root
```

**步骤2** 执行如下命令将镜像启动默认用户加入root属组。

```
usermod -a -G root 用户名
```

**步骤3** 切换回普通用户。

----结束

## Dockerfile 样例（以昇腾镜像为例）

```
FROM npu_serving:v1

COPY ./requirements-pytorch.txt /home/HwHiAiUser/serving-requirements.txt

RUN pip3 install pip -U -i https://pypi.tuna.tsinghua.edu.cn/simple --trusted-host pypi.tuna.tsinghua.edu.cn && \
    pip3 install -r serving-requirements.txt -i https://pypi.tuna.tsinghua.edu.cn/simple --trusted-host
    pypi.tuna.tsinghua.edu.cn

#确保Python软连接为Python3，若Python已指向Python3，则无需配置
RUN mv /usr/bin/python /usr/bin/python_bak
    ln -s /usr/local/python3.7.5/bin/python3 /usr/bin/python

#以下步骤为普镜像启动用户通用户为时需要的额外配置,以昇腾镜像为例(昇腾镜像默认用户为HwHiAiUser)
USER root
#将普通用户加入root属组
RUN usermod -a -G root HwHiAiUser
#切换为普通用户
USER HwHiAiUser
```

### 3.1.7.3 在线服务

#### 3.1.7.3.1 查看在线服务列表

在页面中可以查看ID、服务名称、框架类型、服务描述、状态、运行节点数/总节点数、调用失败次数/总次数、服务类型、节点组、创建时间、操作（编辑、启动、停止、删除、预测、回滚）等信息。

## 说明

- “运行中”状态的任务，可进行“停止”、“更多>预测”操作。
- “已停止”状态的任务，可进行“编辑”、“启动”、“删除”、“更多>回滚”操作。
- “部署中”状态的任务，可进行“停止”操作。
- “运行失败”状态的任务，可进行“编辑”、“启动”、“删除”、“更多>回滚”操作。

## 操作步骤

**步骤1** 选择“推理服务>在线服务”。

进入“在线服务”界面。

**图 3-93 在线服务列表**

The screenshot shows a table with one row of data. The columns include ID, Service Name, Frame Type, Service Description, Status, Running Node Count/Total Node Count, Failed Call Count/Total Call Count, Service Type, and Node Group. The service 'yolo1' is listed with a status of '已停止' (Stopped), 0/1 running nodes, and 0/0 failed calls. There are buttons for Edit, Start, Delete, and More.

**步骤2** 单击任意一条服务名称，查看在线服务详情。

**图 3-94 在线服务详情**

The screenshot shows the deployment details for service 'yolo1'. It includes fields for Name (yolo1), Frame Type (Platform), Type (HTTP mode), Status (Stopped), Running Node Count/Total Node Count (0/1), Failed Call Count/Total Call Count (0/0), and Model Configuration (model-V0001). Below this, there are tabs for Call Guide, Prediction, Monitoring Information, Log, and Deployment Record. The 'Call Guide' tab is selected, showing the API endpoint: https://90.90.161.166/serving-gateway/968f4ea3e6b41b1bac8837a738810e/infer. There is also a 'Copy' button for the URL. A 'Parameter Configuration' section is shown below.

----结束

### 3.1.7.3.2 创建在线服务

#### 前提条件

数据准备：在创建在线服务之前，请确保已经准备好可用的模型、推理脚本、镜像等供新建服务使用。详情参见：

- [上传模型](#)
- [3.1.5.2 训练任务](#)
- [上传算法](#)
- [3.1.8 镜像管理](#)

#### 操作步骤

**步骤1** 选择“推理服务 > 在线服务”。

进入“在线服务”页面。

**步骤2** 单击“创建”。

进入“创建在线服务”页面。

**图 3-95** 创建在线服务

**步骤3** 输入基本信息。

**表 3-26** 参数说明

| 参数   | 说明  |
|------|---|
| 服务名称 | 填写在线服务名称。<br>支持字母、数字、汉字、英文横杠和下划线，在线服务名称长度不能超出32个字符，不可重复（ID为自动递增）。 |

| 参数    | 说明  |
|-------|---|
| 框架类型  | <ul style="list-style-type: none"> <li>平台</li> <li>自定义</li> </ul>   |
| 服务类型  | <ul style="list-style-type: none"> <li>HTTP模式（若用户使用自定义框架类型，服务类型只支持HTTP模式）</li> <li>GRPC模式</li> </ul>  |
| 服务描述  | 任务描述，字符长度不能超出200。   |
| 请求类型  | <p>（框架类型为自定义时选择）</p> <ul style="list-style-type: none"> <li>POST</li> <li>GET</li> </ul>  |
| 服务端口号 | <p>（框架类型为自定义时填写）</p> <p>输入服务端口号。端口号最大不超过65565。输入的服务端口号需要与用户实际对外提供访问的服务端口一致。</p>   |
| API   | <p>（框架类型为自定义时填写）</p> <p>用户输入自定义的推理服务API接口地址。</p> <p>必须以“/”为开头，只支持大小写英文字母、数字和特殊字符-_.~/，最多128个字符。</p> <p>输入的API接口地址需要与用户实际对外提供访问的API接口地址一致。</p> |
| 节点组   | 从下拉列表中选择在线服务关联的节点组。   |

**步骤4** 根据框架类型填写在线服务参数。

**表 3-27 平台框架服务参数说明**

| 参数   | 说明  |
|------|---|
| 模型类型 | <ul style="list-style-type: none"> <li>无</li> <li>我的模型</li> <li>预训练模型</li> </ul>                      |
| 模型   | <p>（模型选择“我的模型”或“预训练模型”时选择）选择部署在线服务时所需的模型和模型版本。</p> <p>选择模型后，容器内模型挂载路径为/usr/local/serving/models/。</p> |

| 参数         | 说明   |
|------------|--|
| 自定义推理脚本    | <p>选择或单击“这里”上传推理脚本，进入上传页面。</p> <p><b>推理脚本模板如下：</b></p> <pre>import os import torch from logger import Logger  # 平台框架日志库，日志等级有(debug, info, warning, error, critical) 具体用法： log.info("xxxx") log = Logger().logger  # 只能定义一个class class CommonInferenceService:     # 请在__init__初始化方法中接收args参数，并加载模型（其中模型路径参数为args.model_path，是否使用gpu参数为args.use_gpu，模型加载方法用户可自定义）     def __init__(self, args):         # 相关参数         self.args = args         # 初始化加载模型         self.model = self.load_model()      # 读取模型，初始化读取模型，做为类初始化时加载模型至内存，模型路径均在self.args.model_path中，     # 用户可自主修改，以下为pytorch模型例子：     def load_model(self):         log.info("===== &gt; load model &lt; =====")         if os.path.isfile(self.args.model_path):             self.checkpoint = torch.load(self.args.model_path)         else:             for file in os.listdir(self.args.model_path):                 self.checkpoint = torch.load(self.args.model_path + file)         model = self.checkpoint["model"]         model.load_state_dict(self.checkpoint['state_dict'])         for parameter in model.parameters():             parameter.requires_grad = False         if self.args.use_gpu:             model.cuda()         model.eval()         return model      """ 注意：inference方法名称固定此方法名必须存在 注意：该方法入参也必须为data，data的类型为dict，即为json，具体内容为 {     "data_path": "/xxx/xxx/xxx.jpg",     "data_name": "aaa.jpg" } 其中data_path为用户上传完的文件保存路径，框架会重命名，避免重名覆盖 data_name为用户上传完的原文件名称     """      def inference(self, data):         # return 推理结果         pass</pre> <p><b>说明</b><br/>使用自定义脚本时，需要符合模板中注释的规则。</p> |
| 灰度发布分流 (%) | 单版本默认100%，多版本可根据适用场景调配百分比(%)，整体满足100%即可。   |

| 参数        | 说明   |
|-----------|--|
| 运行参数模式    | 选择运行参数模式。 <ul style="list-style-type: none"><li>● key-value</li><li>● arguments</li></ul>  |
| 运行参数      | 单击  或  可增加或删除运行参数。   |
| 镜像选择      | 选择部署在线服务时所需的镜像和镜像版本。   |
| 计算类型      | 从下拉列表选择需要的计算节点类型，包括： <ul style="list-style-type: none"><li>● 基础资源</li><li>● GPU / 整卡</li><li>● GPU / MIG</li><li>● NPU / 整卡</li><li>● NPU / vNPU</li></ul> |
| 基础资源      | 从下拉框选择基础资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。选择完成后，将显示该基础资源规格参数信息，包括CPU个数、内存大小。   |
| GPU/NPU规格 | （当计算类型为GPU或NPU时需要选择）从下拉列表中选择在线服务时所需的资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。选择完成后，将显示该资源规格参数。  |
| 节点数量      | 默认为1，多版本可根据适用场景调配节点数，整体配置最大值不超出10节点即可。   |

表 3-28 自定义框架服务参数说明

| 参数     | 说明   |
|--------|--|
| 模型类型   | <ul style="list-style-type: none"><li>● 无</li><li>● 我的模型</li><li>● 预训练模型</li></ul>         |
| 模型     | （模型选择“我的模型”或“预训练模型”时选择）选择部署在线服务时所需的模型和模型版本。<br>选择模型后，容器内模型挂载路径为/usr/local/serving/models/。 |
| 推理服务脚本 | 选择推理服务脚本。<br>选择后，容器内脚本挂载路径为/usr/local/serving/algorithm/。                                  |
| 运行命令   | 输入运行命令。最多支持输入8192个字符。  |

| 参数        | 说明   |
|-----------|--|
| 灰度发布分流(%) | 单版本默认100%，多版本可根据适用场景调配百分比(%)，整体满足100%即可。   |
| 镜像选择      | 选择部署在线服务时所需的镜像和镜像版本。   |
| 计算类型      | 从下拉列表选择需要的计算节点类型，包括： <ul style="list-style-type: none"><li>• 基础资源</li><li>• GPU / 整卡</li><li>• GPU / MIG</li><li>• NPU / 整卡</li><li>• NPU / vNPU</li></ul> |
| 基础资源      | 从下拉框选择基础资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。选择完成后，将显示该基础资源规格参数信息，包括CPU个数、内存大小。   |
| GPU/NPU规格 | (当计算类型为GPU或NPU时需要选择)从下拉列表中选择在线服务时所需的资源规格（资源规格可由超级管理员、组织管理员在控制台 > 资源规格管理界面进行创建、修改）。选择完成后，将显示该资源规格参数。  |
| 节点数量      | 默认为1，多版本可根据适用场景调配节点数，整体配置最大值不超出10节点即可。   |

## 步骤5 配置调度器参数。

表 3-29 调度器配置

| 参数       | 说明   |
|----------|--|
| 调度器类型    | AI Space的调度器类型，超级管理员用户可在控制台 > 调度配置界面修改。<br>当调度器类型为Volcano时，填写下方调度参数。 |
| 所属队列     | 从下拉列表中选择在线服务任务所属队列。  |
| 优先级      | 选择在线服务任务的优先级。<br>当集群中运行多个job时，Volcano将以用户定义的优先级调度资源。                 |
| 最少运行pod数 | 设置在线服务任务最少运行pod数，最少运行pod数需小于节点数。                                     |
| 最大重启次数   | 设置在线服务任务最大重启次数。  |
| 生命周期策略   | 选择Pod生命周期策略。支持“pod驱逐重启作业”，当pod被驱逐时，将重启该job。                          |

| 参数    |     | 说明   |
|-------|-----|--|
| 节点亲和性 | 属性  | 输入节点属性，即节点标签，设置节点亲和性前需要为节点打上对应的标签。<br>单击  或  可添加或删除属性。   |
|       | 类型  | 从下拉列表中选择节点亲和性类型，包括硬亲和性调度策略、软亲和性调度策略。 <ul style="list-style-type: none"> <li>硬亲和性调度策略：调度器必须满足，多条规则间是一种“或”的关系，即只需要满足一条规则即会进行调度。</li> <li>软亲和性调度策略：调度器会尽量满足，无论是满足其中一条或者是都不满足都会进行调度。</li> </ul>                            |
|       | 操作符 | 从下拉列表中选择操作符。 <ul style="list-style-type: none"> <li>In：标签的值在某个列表中</li> <li>NotIn：标签的值不在某个列表中</li> <li>Exists：某个标签存在</li> <li>DoesNotExist：某个标签不存在</li> <li>Gt：标签的值大于某个值（字符串比较）</li> <li>Lt：标签的值小于某个值（字符串比较）</li> </ul> |
|       | 取值  | 添加属性取值。<br>通过配置节点亲和性规则，调度器可以将Pod调度到具有特定标签的节点。  |

**步骤6**（可选）若用户需要进行灰度发布，单击“添加分流进行灰度发布”，填写服务参数信息。最多支持2个版本。

所有版本的灰度发布分流之和应为100%。

**步骤7** 单击“提交”。

跳转至在线服务列表页面且该条新建任务状态变为“部署中”。

**图 3-96 创建在线服务成功**



## 说明书

服务状态转变为运行中说明容器已启动成功，但模型加载需要一定时间，因此部署完成的模型需一段时间后才能成功预测，请耐心等待。

出现运行失败可能的原因如下：

- 配置的节点数或规格资源不足。
- 参数配置、模型错误。
- 模型与推理脚本不匹配。
- 镜像错误。

----结束

## 自定义框架在线服务镜像示例

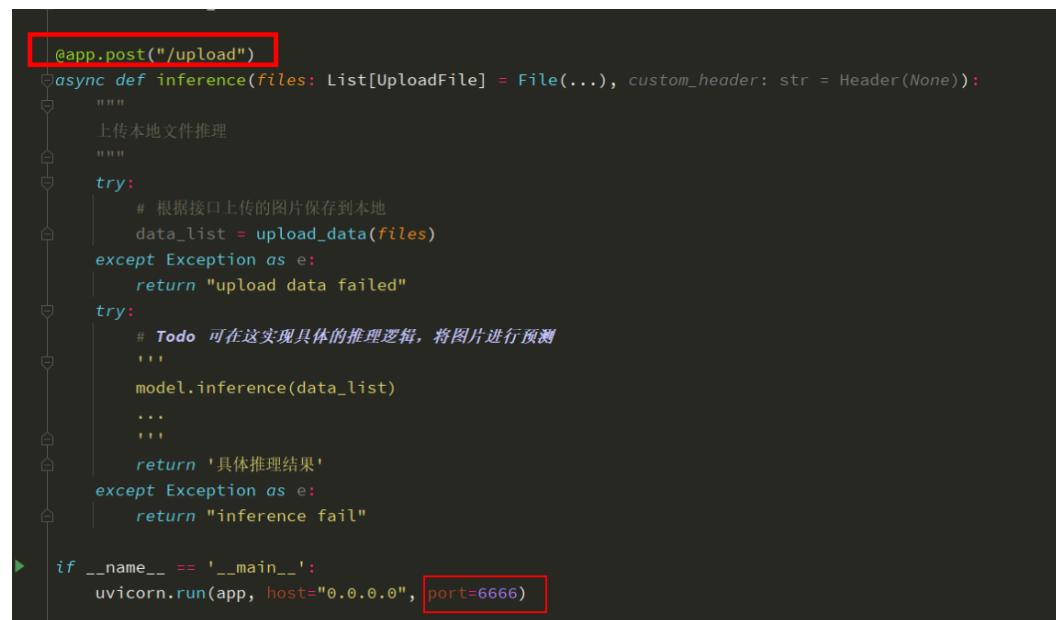
以Python的fastAPI Web框架构建简易自定义镜像为例，Demo的内容如下：

```
#!/usr/bin/python
# -*- coding: utf-8 -*-
from tempfile import NamedTemporaryFile
from typing import List
import os
from pathlib import Path
import shutil
from fastapi import FastAPI, File, UploadFile, Header
import uvicorn
app = FastAPI(version='1.0', title="Zhejiang Lab TS_Serving inference Automation",
description="API for performing oneflow、tensorflow、pytorch inference")
def upload_data(files):
"""
    前端上传图片保存到本地
"""
    save_data_dir = "/tmp/data/"
    if not os.path.exists(save_data_dir):
        os.mkdir(save_data_dir)
    data_list = list()
    for file in files:
        try:
            suffix = Path(file.filename).suffix
            with NamedTemporaryFile(delete=False, suffix=suffix, dir=save_data_dir) as tmp:
                shutil.copyfileobj(file.file, tmp)
                tmp_file_name = Path(tmp.name).name
                data = {"data_name": file.filename, "data_path": save_data_dir + tmp_file_name}
                data_list.append(data)
        finally:
            file.file.close()
    return data_list
@app.post("/upload")
async def inference(files: List[UploadFile] = File(...), custom_header: str = Header(None)):
"""
    上传本地文件推理
"""
    try:
        # 根据接口上传的图片保存到本地
        data_list = upload_data(files)
    except Exception as e:
        return "upload data failed"
    try:
        # Todo 可在这实现具体的推理逻辑，将图片进行预测
        """
            model.inference(data_list)
            ...
        """
    except:
        return "specific inference result"
    except Exception as e:
        return "inference fail"
```

```
if __name__ == '__main__':
    uvicorn.run(app, host="0.0.0.0", port=6666)
```

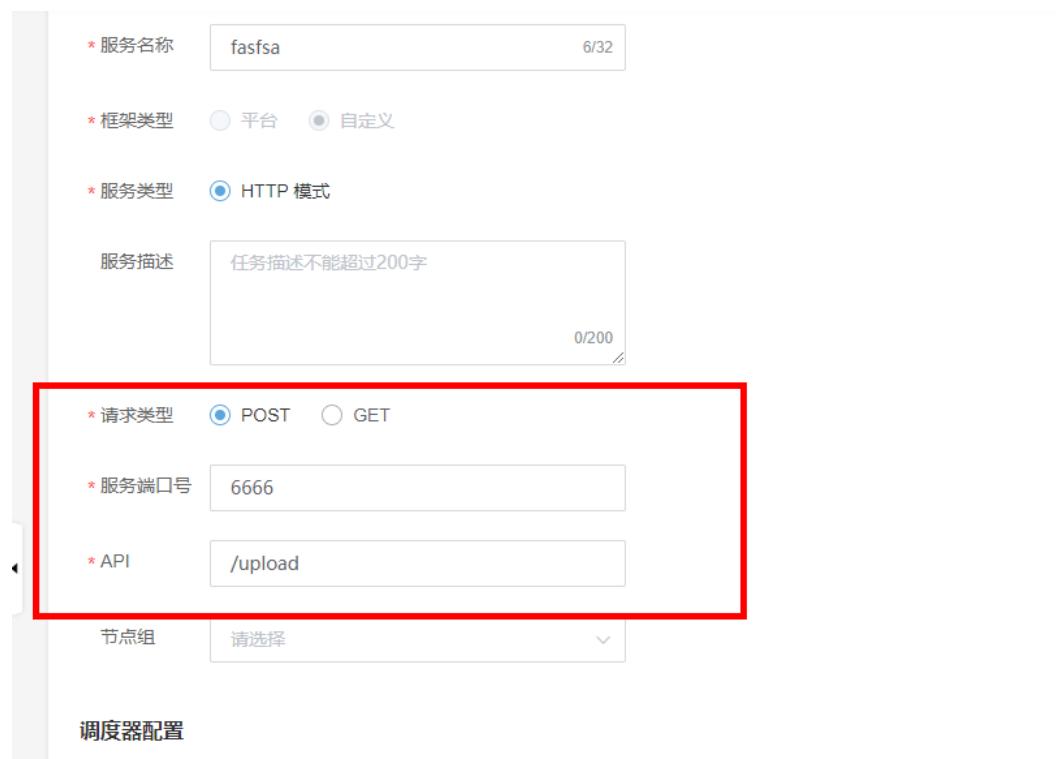
创建自定义框架在线推理服务时，填写的服务端口号和API路径需与镜像中自定义内容一致。

图 3-97 示例



```
@app.post("/upload")
async def inference(files: List[UploadFile] = File(...), custom_header: str = Header(None)):
    """
    上传本地文件推理
    """
    try:
        # 根据接口上传的图片保存到本地
        data_list = upload_data(files)
    except Exception as e:
        return "upload data failed"
    try:
        # Todo 可在这实现具体的推理逻辑，将图片进行预测
        ...
        model.inference(data_list)
        ...
    except Exception as e:
        return "inference fail"

if __name__ == '__main__':
    uvicorn.run(app, host="0.0.0.0", port=6666)
```



### 3.1.7.3.3 启停在线服务

用户可在“在线服务”列表或部署详情页面对在线服务信息进行启动、停止操作。

## 操作步骤

**步骤1** 启动状态为“已停止”、“运行失败”的在线推理服务。

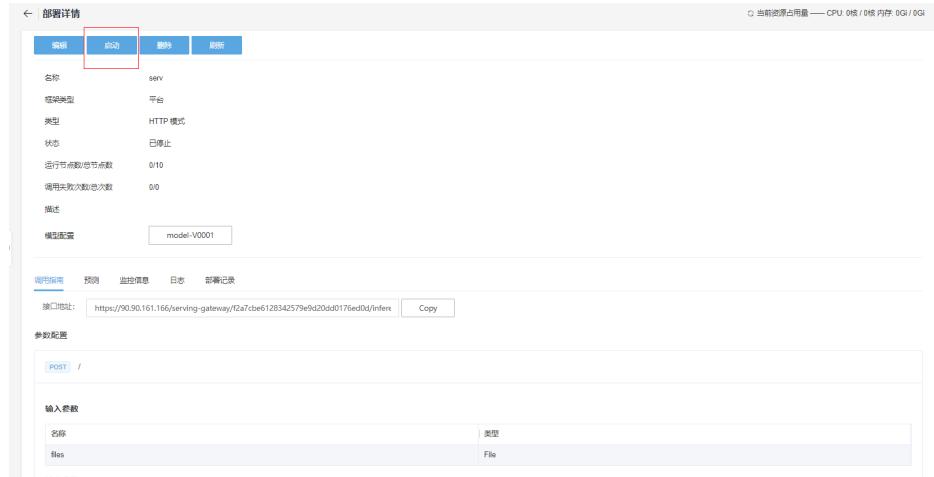
- 单击“推理服务 > 在线服务”，进入“在线服务”列表页面单击“启动”进行启动服务。

图 3-98 启动在线服务



- 单击“在线服务”列表页服务名称进入“部署详情”页面，单击“启动”进行启动服务。

图 3-99 启动在线服务



**步骤2** 停止状态为“部署中”、“运行中”的在线推理服务。

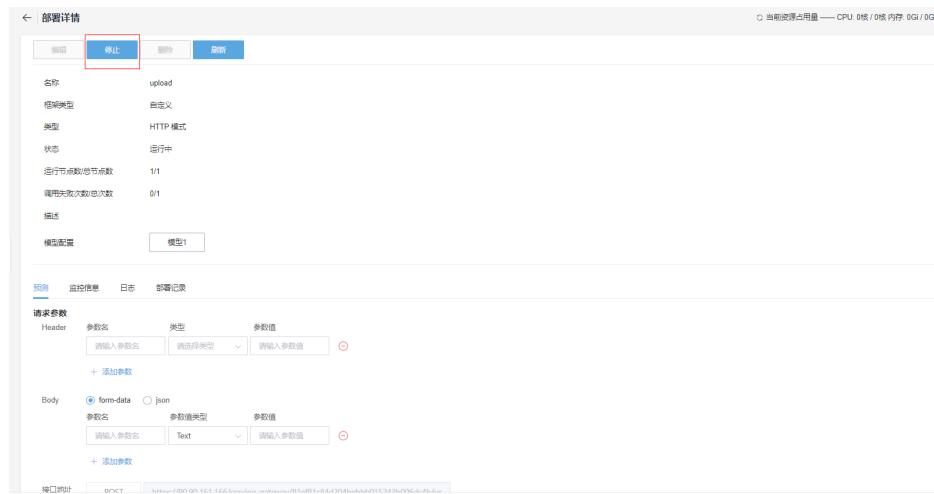
- 单击“推理服务 > 在线服务”，进入“在线服务”列表页面单击“停止”进行停止服务。

图 3-100 停止在线服务



- 单击“在线服务”列表页服务名称进入“部署详情”页面，单击“停止”按钮进行停止服务。

图 3-101 停止在线服务



----结束

### 3.1.7.3.4 预测在线服务

用户可在“在线服务”列表或部署详情页面，对状态为“运行中”的在线服务进行预测操作。

## 预测平台框架服务

**步骤1** 选择“推理服务 > 在线服务”。

进入“在线服务”页面。

**步骤2** 选择“更多 > 预测”。

进入“预测”列表页面。

图 3-102 在线服务列表

| ID | 服务名称 | 框架类型 | 服务描述 | 状态  | 运行节点数/总节点数 | 调用失败次数/总次数 | 服务类型    | 节点组 | 操作  |
|----|------|------|------|-----|------------|------------|---------|-----|---|
| 2  | 1233 | 自定义  |      | 运行中 | 1/1        | 0/1        | HTTP 模式 |     | <a href="#">编辑</a> <a href="#">停止</a> <a href="#">删除</a> <a href="#">更多</a> |
| 1  | serv | 平台   |      | 已停止 | 0/1        | 0/0        | HTTP 模式 |     | <a href="#">编辑</a> <a href="#">启动</a> <a href="#">删除</a> <a href="#">预测</a> |

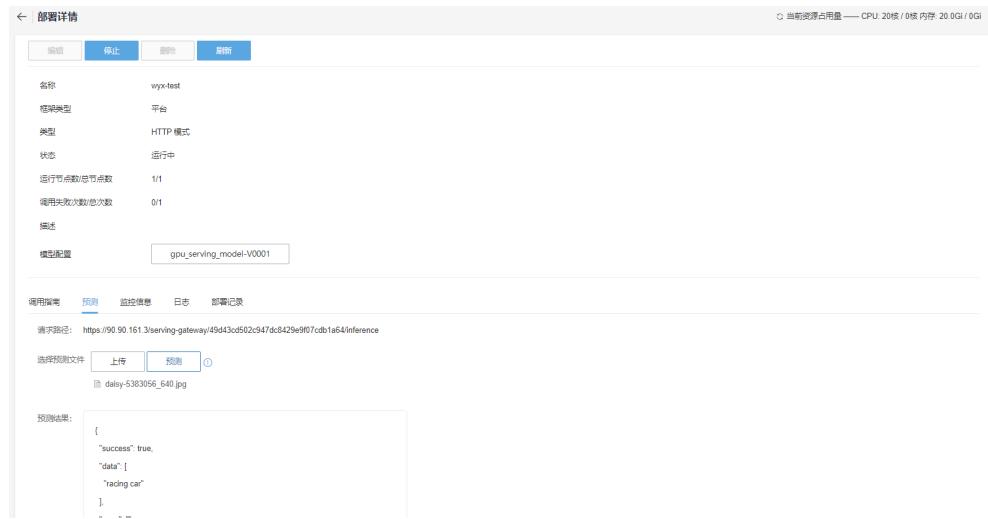
### 说明

您也可以单击服务名称进入“部署详情”页面，单击“预测”进入“预测”列表页面。

**步骤3** 在线服务预测。

1. 单击“上传”，进行上传文件。
2. 文件上传完成后，单击“预测”，进行在线服务预测。

图 3-103 预测在线服务



### 说明

仅支持预测.JPG、.JPEG、.PNG、.BMP格式的文件，且单次预测选择的文件大小总计不超过10MB。

----结束

## 预测自定义框架服务

**步骤1** 选择“推理服务 > 在线服务”。

进入“在线服务”页面。

**步骤2** 选择“更多 > 预测”。

进入“预测”列表页面。

图 3-104 在线服务列表

| ID | 服务名称 | 框架类型 | 服务描述 | 状态    | 运行节点数/总节点数 | 调用失败次数/总次数 | 服务类型    | 节点组 | 操作            |
|----|------|------|------|-------|------------|------------|---------|-----|---------------|
| 2  | 1233 | 自定义  |      | ● 运行中 | 1/1        | 0/1        | HTTP 模式 |     | 编辑 停止 删除 更多 ▾ |
| 1  | serv | 平台   |      | ● 已停止 | 0/1        | 0/0        | HTTP 模式 |     | 编辑 启动 删除 预测   |

### 说明

您也可以单击服务名称进入“部署详情”页面，单击“预测”进入“预测”列表页面。

**步骤3** 在线服务预测。

1. 选择“预测”页签。
2. 填写请求参数。

表 3-30 请求参数

| 参数     | 说明  |
|--------|---|
| Header | 输入参数名、类型、参数值。<br>单击  添加参数。 |
| Body   | 选择上传模式，包括form-data , json后输入参数名、参数值类型、参数值。  |
| 接口地址   | 接口地址默认不可修改。   |

3. 单击“预测”。

**步骤4** 预测的结果将展示在“预测结果”框内。

----结束

### 3.1.7.3.5 编辑在线服务

用户可在“在线服务”列表或部署详情页面对在线服务信息进行编辑。

#### 操作步骤

**步骤1** 选择“推理服务 > 在线服务”。

进入“在线服务”页面。

**步骤2** 在待编辑服务名称所在行，单击“编辑”。

进入“编辑在线服务”页面。

图 3-105 编辑在线服务



#### 说明

您也可以单击服务名称进入“部署详情”页面，单击“编辑”进入“编辑在线服务”页面。

**步骤3** 根据**3.1.7.3.2 创建在线服务**对参数信息进行编辑。

**步骤4** 单击“提交”，跳转至在线服务列表页面且该条任务状态变为“部署中”。

图 3-106 编辑在线服务成功



----结束

### 3.1.7.3.6 回滚在线服务

用户可在“在线服务”列表或部署详情页面对状态为“已停止”或“运行失败”的在线服务进行回滚。

#### 操作步骤

- 步骤1** 选择“推理服务 > 在线服务”。
- 进入“在线服务”页面。
- 步骤2** 在需要回滚服务名称所在行，单击“回滚”。
- 进入“部署详情”页面。

图 3-107 回滚在线服务入口

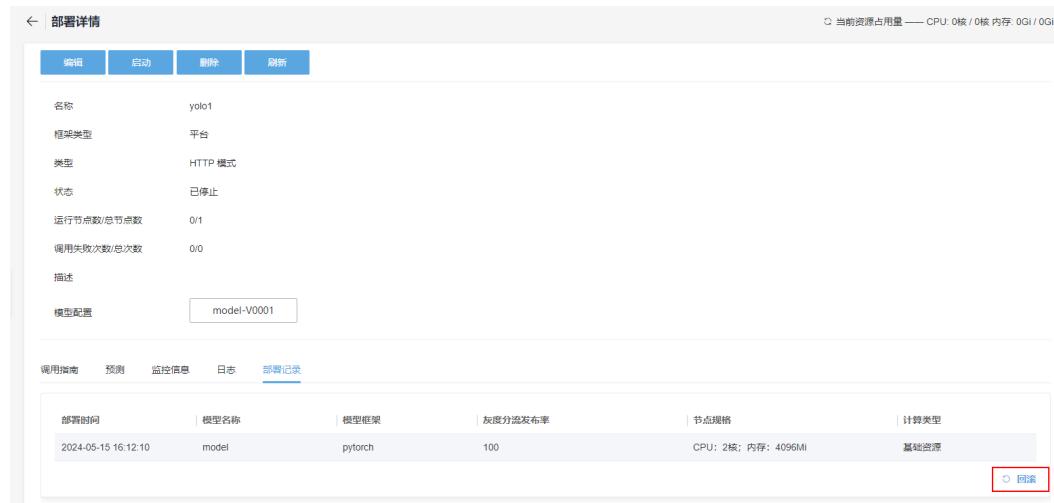


#### 说明

您也可以单击服务名称进入“部署详情”页面，选择“部署记录”进入后续操作。

- 步骤3** 在“部署记录”页签，选择需要的版本，单击“回滚”。

图 3-108 回滚在线服务



**步骤4** 单击“确定”，返回在线服务列表页，此时该任务状态为“部署中”。

----结束

### 3.1.7.3.7 删除在线服务

用户可在“在线服务”列表或部署详情页面，对状态为“已停止”“运行失败”的在线服务信息进行删除。

## 操作步骤

**步骤1** 选择“推理服务 > 在线服务”。

进入“在线服务”页面。

图 3-109 在线服务列表

| ID | 服务名称  | 框架类型 | 服务描述 | 状态  | 运行节点数/总节点数 | 调用失败次数/总次数 | 服务类型    | 节点组 | 操作  |
|----|-------|------|------|-----|------------|------------|---------|-----|---|
| 1  | yolo1 | 平台   |      | 已停止 | 0/1        | 0/0        | HTTP 模式 |     | <a href="#">编辑</a> <a href="#">启动</a> <a href="#">删除</a> <a href="#">更多</a> |

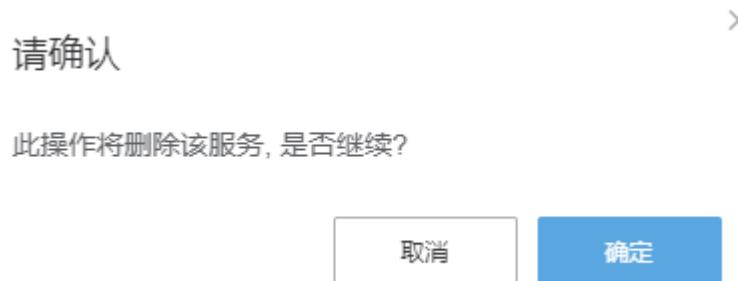
## 说明

您也可以单击服务名称进入“部署详情”页面后进行后续操作。

**步骤2** 在待删除的在线服务所在行，单击“删除”。

弹出提示框。

图 3-110 删除在线服务



**步骤3** 单击“确定”，删除当前服务。

#### 须知

该操作将会删除该任务的所有信息且删除后无法恢复，请谨慎操作。

#### ----结束

### 3.1.7.4 批量服务

#### 3.1.7.4.1 查看批量服务列表

在页面中可以查看到ID、服务名称、框架类型、服务描述、状态、进度、任务开始时间、节点组、任务结束时间、操作（Fork、重新推理、停止、删除、结果下载）等信息。

##### 说明

- “运行中”状态的任务，可进行“停止”等操作。
- “部署中”状态的任务，可进行“停止”等操作。
- “已停止”状态的任务，可进行“重新推理”、“删除”等操作。
- “运行失败”状态的任务，可进行“重新推理”、“删除”等操作。
- “运行完成”状态的任务，可进行“Fork”、“删除”、“结果下载”等操作。
- “未知”状态的任务，可进行“删除”等操作。

### 操作步骤

**步骤1** 选择“推理服务 > 批量服务”。

进入“批量服务”页面。

图 3-111 批量服务列表

| 批量服务                |      |      |      |       |    |        |              |        |                 |  |
|---------------------|------|------|------|-------|----|--------|--------------|--------|-----------------|--|
| 创建                  |      |      |      |       |    |        |              |        |                 |  |
| ID                  | 服务名称 | 框架类型 | 服务描述 | 状态    | 进度 | 任务开始时间 | 节点组          | 任务结束时间 | 操作              |  |
| 1                   | 平台   |      |      | ● 已停止 | 0% | --     | node_group_1 | --     | 编辑 重新推理 删除 结果下载 |  |
| 共 1 条 10 条 < 1 > 跳转 |      |      |      |       |    |        |              |        |                 |  |

**步骤2** 单击列表中任意一条批量服务名称，进入部署详情列表页。

可查看具体的批量服务信息和日志等板块。

**图 3-112 部署详情列表**

The screenshot shows the 'Deployment Details' page for a service named 'test'. Key details include:

- 名称:** test
- 状态:** 运行完成
- 镜像信息:** a800-image:v1
- 推理脚本:** gpu\_serving
- 计算类型:** 基础推理
- 基础资源:** CPU: 10核; 内存: 10240Mi
- 节点数量:** 1
- 描述:** (empty)
- 进度:** 100%
- 模型名称&版本:** gpu\_serving\_model-V0001-pytorch
- 任务开始时间:** 2024-05-07 14:26:44
- 任务结束时间:** 2024-05-07 14:26:52
- 输入数据目录:** serving/Input/7/2024050714222692u8Qi
- 结束下载:** 下载

The 'Logs' section shows AI Service Log entries:

```
[AI Service Log] 2024-05-07T14:26:43.583100000+08:00 stderr F /bin/bash: jq: command not found
[AI Service Log] 2024-05-07T14:26:43.585138342+08:00 stderr F cat: /home/hostfile.json: No such file or directory
[AI Service Log] 2024-05-07T14:26:47.484595819+08:00 stdout F ResNet(
[AI Service Log] 2024-05-07T14:26:47.484627207+08:00 stdout F  (conv1): Conv2d(3, 64, kernel_size=(7, 7), stride=(2, 2), padding=(3, 3), bias=False)
[AI Service Log] 2024-05-07T14:26:47.484632619+08:00 stdout F  (bn1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
[AI Service Log] 2024-05-07T14:26:47.484632619+08:00 stdout F  (relu1): ReLU(inplace=True)
```

----结束

### 3.1.7.4.2 创建批量服务

#### 前提条件

数据准备：在创建批量服务之前，请确保已经准备好可用的模型、推理脚本、镜像等供新建服务使用。

详情参见：

- [上传模型](#)
- [3.1.5.2 训练任务](#)
- [上传算法](#)
- [3.1.8 镜像管理](#)

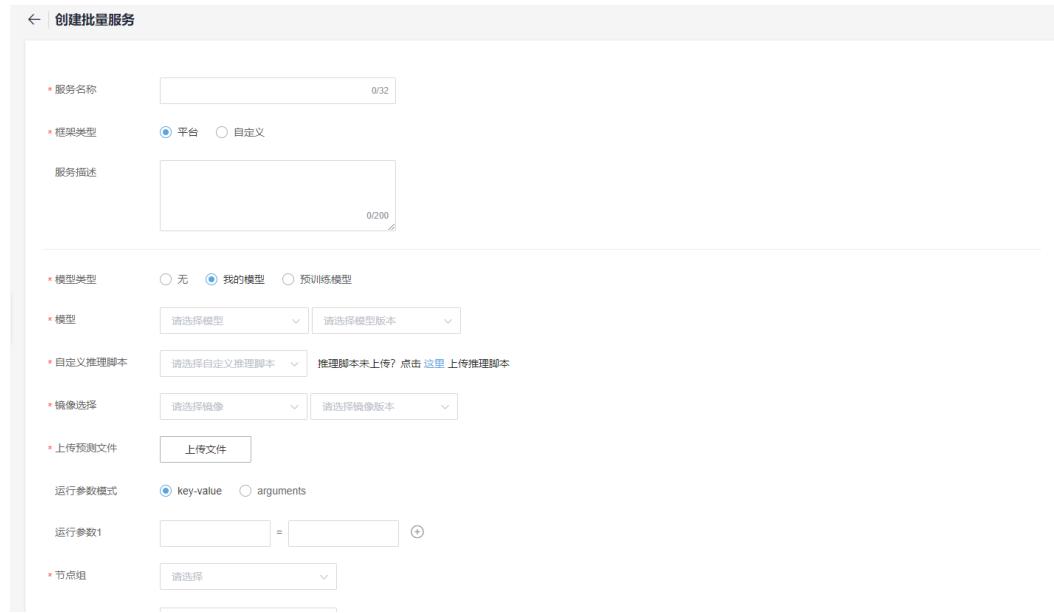
#### 操作步骤

**步骤1** 选择“推理服务 > 批量服务”。

进入“批量服务”页面。

**步骤2** 单击“创建”。

弹出“创建批量服务”页面。

**图 3-113 创建批量服务****步骤3 填写基本信息。****表 3-31 参数说明**

| 参数   | 说明   |
|------|--|
| 服务名称 | 填写服务名称。<br>支持字母、数字、汉字、英文横杠和下划线、字符长度不能超出32，服务名称不可重复（ID为自动递增）。   |
| 框架类型 | <ul style="list-style-type: none"><li>平台</li><li>自定义</li></ul> |
| 服务描述 | 任务描述，字符长度不能超出200。  |

**步骤4 根据框架类型填写推理服务参数。****表 3-32 平台框架参数说明**

| 参数   | 说明  |
|------|---|
| 模型类型 | 选择模型类型。 <ul style="list-style-type: none"><li>无</li><li>我的模型</li><li>预训练模型</li></ul>              |
| 模型   | (当模型类型选择“我的模型”或“预训练模型”时选择)<br>选择创建批量服务时所需的模型、模型版本。<br>选择模型后，容器内模型挂载路径为/usr/local/serving/models/。 |

| 参数        | 说明   |
|-----------|--|
| 自定义推理脚本   | 选择自定义推理脚本，或单击“上传”上传自定义脚本。  |
| 镜像选择      | 选择部署推理服务所需镜像和镜像版本。   |
| 上传预测文件    | 单击“选择”上传预测文件，单个文件大小不超过5MB，最多支持上传100个文件。  |
| 运行参数模式    | 选择运行参数模式。 <ul style="list-style-type: none"><li>• key-value</li><li>• arguments</li></ul>  |
| 运行参数      | 输入运行参数。<br>单击 $\oplus$ 或 $\ominus$ 可增加或删除运行参数。   |
| 节点组       | 选择批量推理服务关联的节点组。  |
| 计算类型      | 从下拉列表中选择创建批量服务时所需的计算节点类型，包括： <ul style="list-style-type: none"><li>• 基础资源</li><li>• GPU / 整卡</li><li>• GPU / MIG</li><li>• NPU / 整卡</li><li>• NPU / vNPU</li></ul> |
| 基础资源      | 从下拉框选择基础资源规格（资源规格可由超级管理员、组织管理员在控制台>资源规格管理界面进行创建、修改）。选择完成后，将显示该基础资源规格参数信息，包括CPU个数、内存大小。   |
| GPU/NPU规格 | （当计算类型为GPU或NPU时需要选择）从下拉列表中选择批量服务时所需的资源规格（资源规格可由超级管理员在控制台进行增删改）。选择完成后，将显示该资源规格参数。   |
| 节点数量      | 默认为1，多版本可根据适用场景调配节点数，整体配置最大值不超出10节点即可。   |

表 3-33 自定义框架参数说明

| 参数   | 说明   |
|------|--|
| 模型类型 | 选择模型类型。 <ul style="list-style-type: none"><li>• 无</li><li>• 我的模型</li><li>• 预训练模型</li></ul> |

| 参数        | 说明   |
|-----------|--|
| 模型        | (当模型类型选择“我的模型”或“预训练模型”时选择)<br>选择创建批量服务时所需的模型、模型版本。<br>选择模型后，容器内模型挂载路径为/usr/local/serving/models/。  |
| 运行命令      | 输入运行命令。最多支持输入8192个字符。  |
| 镜像选择      | 选择推理服务镜像和镜像版本。   |
| 节点组       | 从下拉列表中选择批量服务关联的节点组。  |
| 上传预测文件    | 单击“上传文件”上传预测文件，单个文件大小不超过5MB，最多支持上传100个文件。  |
| 文件保存路径    | 用户自定义上传的预测文件保存的路径。<br>路径必须以“/”为开头，只支持大小写英文字母、数字和特殊字符-_.~/，最多128个字符。  |
| 结果保存路径    | 用户自定义推理结果保存的路径。<br>路径必须以“/”为开头，只支持大小写英文字母、数字和特殊字符-_.~/，最多128个字符。   |
| 节点组       | 选择批量推理服务关联的节点组。  |
| 计算类型      | 从下拉列表中选择创建批量服务时所需的计算节点类型，包括： <ul style="list-style-type: none"><li>• 基础资源</li><li>• GPU / 整卡</li><li>• GPU / MIG</li><li>• NPU / 整卡</li><li>• NPU / vNPU</li></ul> |
| 基础资源      | 从下拉框选择基础资源规格（资源规格可由超级管理员、组织管理员在控制台>资源规格管理界面进行创建、修改）。<br>选择完成后，将显示该基础资源规格参数信息，包括CPU个数、内存大小。   |
| GPU/NPU规格 | (当计算类型为GPU或NPU时需要选择)从下拉列表中选择批量服务时所需的资源规格（资源规格可由超级管理员在控制台进行增删改）。<br>选择完成后，将显示该资源规格参数。   |
| 节点数量      | 默认为1，整体配置最大值不超出10节点即可。   |

## 步骤5 配置调度器参数。

表 3-34 调度器配置

| 参数       | 说明  |
|----------|---|
| 调度器类型    | AI Space的调度器类型，超级管理员用户可在控制台 > 调度配置界面修改。<br>当调度器类型为Volcano时，填写下方调度参数。  |
| 所属队列     | 从下拉列表中选择批量服务任务所属队列。   |
| 优先级      | 选择批量服务任务的优先级。<br>当集群中运行多个job时，Volcano将以用户定义的优先级调度资源。  |
| 最少运行pod数 | 设置批量服务任务最少运行pod数，最少运行pod数需小于节点数。  |
| 最大重启次数   | 设置批量服务任务最大重启次数。   |
| 生命周期策略   | 选择Pod生命周期策略。支持“pod驱逐重启作业”，当pod被驱逐时，将重启该job。   |
| 节点亲和性    | <p>输入节点属性，即节点标签，设置节点亲和性前需要为节点打上对应的标签。</p> <p>单击  或  可添加或删除属性。</p> |
| 类型       | <p>从下拉列表中选择节点亲和性类型，包括硬亲和性调度策略、软亲和性调度策略。</p> <ul style="list-style-type: none"><li>硬亲和性调度策略：调度器必须满足，多条规则间是一种“或”的关系，即只需要满足一条规则即会进行调度。</li><li>软亲和性调度策略：调度器会尽量满足，无论是满足其中一条或者是都不满足都会进行调度。</li></ul>                                       |
| 操作符      | <p>从下拉列表中选择操作符。</p> <ul style="list-style-type: none"><li>In：标签的值在某个列表中</li><li>NotIn：标签的值不在某个列表中</li><li>Exists：某个标签存在</li><li>DoesNotExist：某个标签不存在</li><li>Gt：标签的值大于某个值（字符串比较）</li><li>Lt：标签的值小于某个值（字符串比较）</li></ul>                |
| 取值       | 添加属性取值。<br>通过配置节点亲和性规则，调度器可以将Pod调度到具有特定标签的节点。   |

**步骤6 单击“提交”。**

跳转至批量服务列表页面，且该条新建任务状态变为“部署中”。

## 说明书

出现运行失败可能的原因如下：

- 新建/编辑/Fork时，配置的节点数或规格资源不足。
- 新建/编辑/Fork时，参数配置、模型错误。
- 新建/编辑/Fork时，模型与推理脚本不匹配。
- 新建/编辑/Fork时，镜像错误。

----结束

### 3.1.7.4.3 启停批量服务

用户可在“批量服务”列表或部署详情页面对批量推理服务信息进行重新推理、停止推理操作。

## 操作步骤

**步骤1** 重新启动状态为“已停止”、“运行失败”的服务。

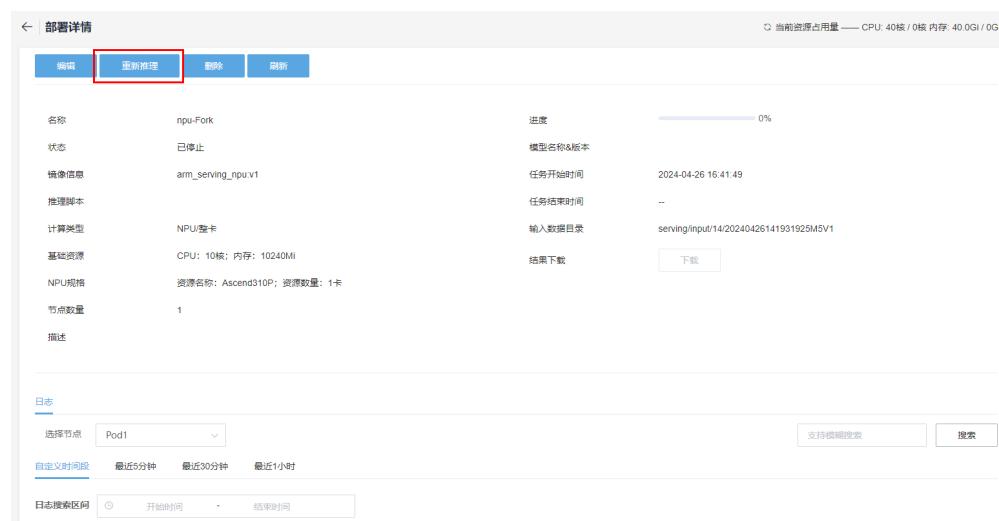
- 选择“推理服务 > 批量服务”，进入“批量服务”列表页面。  
单击“重新推理”启动批量服务。

图 3-114 启动批量服务



- 单击服务名称进入“部署详情”页面，单击“重新推理”启动批量服务。

图 3-115 启动批量服务



**步骤2** 停止状态为“部署中”、“运行中”的服务。

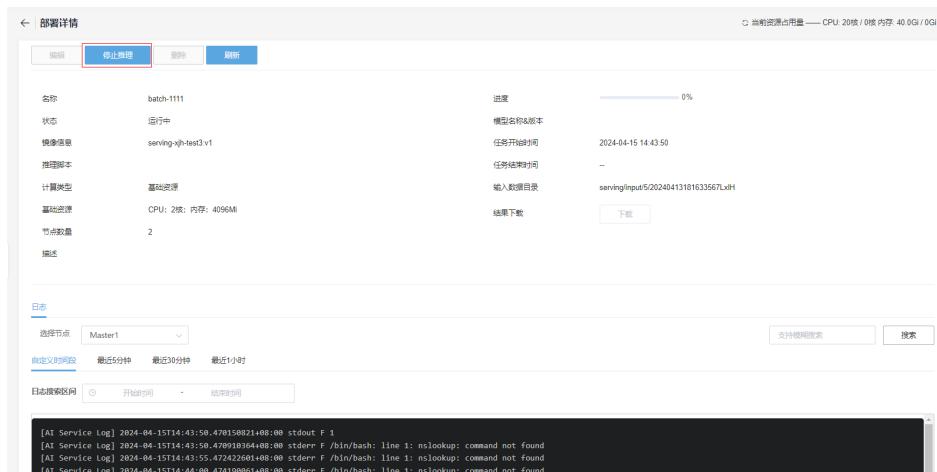
- 选择“推理服务 > 批量服务”，进入“批量服务”列表页面。  
单击“停止推理”停止批量服务。

图 3-116 停止批量服务



- 单击服务名称进入“部署详情”页面，单击“停止推理”停止批量服务。

图 3-117 停止批量服务



----结束

#### 3.1.7.4.4 编辑批量服务

用户可在“批量服务”列表或部署详情页面对状态为“已停止”、“运行失败”的批量推理服务信息进行编辑。

### 操作步骤

**步骤1** 选择“推理服务 > 批量服务”。

进入“批量服务”页面。

#### 说明

您也可以单击服务名称进入“部署详情”页面后进行后续操作。

**步骤2** 在待编辑的批量服务所在行，单击“编辑”。

弹出“编辑批量服务”提示框。

**步骤3** 参见表1，对参数信息进行编辑。

**步骤4** 单击“提交”。

跳转至批量服务列表页面，且该条服务状态变为“部署中”。

## 📖 说明

批量服务新建、编辑后，正常情况下自动运行推理，推理完成后，状态切换为运行完成。

出现运行失败可能的原因如下：

- 配置的节点数或规格资源不足。
- 参数配置、模型错误。
- 模型与推理脚本不匹配。
- 镜像错误。

----结束

### 3.1.7.4.5 Fork 批量服务

用户可在“批量服务”列表或部署详情页面，对状态为“运行完成”的批量推理服务进行Fork。

## 操作步骤

**步骤1** 选择“推理服务 > 批量服务”。

进入“批量服务”列表页面。

**图 3-118 Fork 批量服务入口**

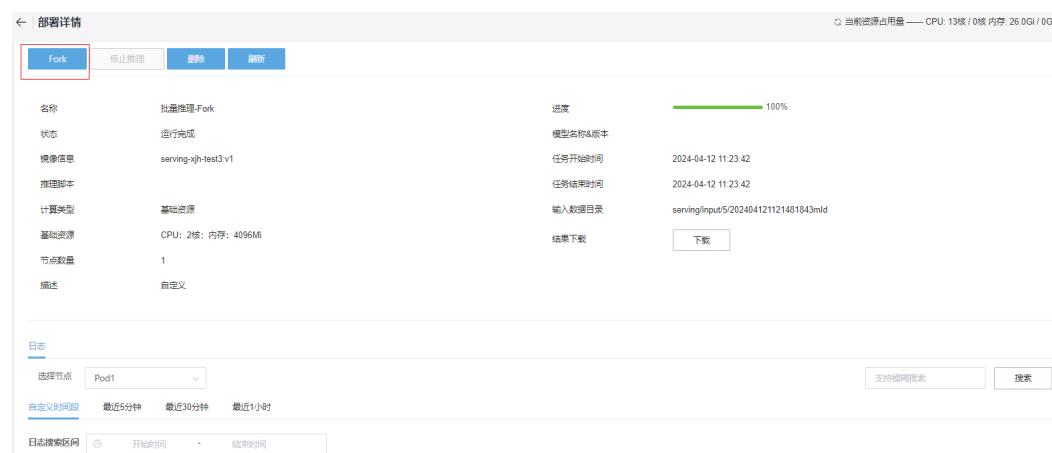


**步骤2** 在待操作的批量服务所在行，单击“Fork”。

弹出“Fork 批量服务”提示框。

## 📖 说明

您也可以单击服务名称进入“部署详情”页面，单击“Fork”。



**步骤3** 参见**3.1.7.4.2 创建批量服务**，对参数信息进行编辑。

**步骤4 单击“提交”。**

返回批量服务列表页面且该条Fork任务状态变为“部署中”。

----结束

### 3.1.7.4.6 删除批量服务

用户可在“批量服务”列表或部署详情页面对批量推理服务进行删除。

#### □ 说明

对于“部署中”、“运行中”状态的任务，不可进行“删除”操作。

### 操作步骤

**步骤1 选择“推理服务 > 批量服务”。**

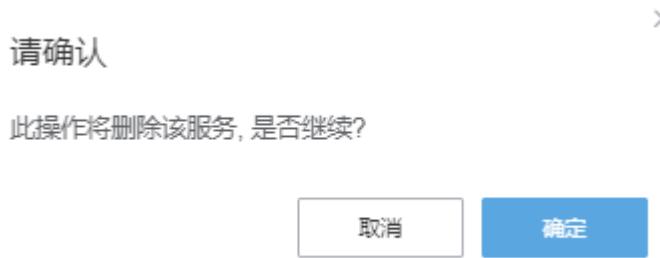
进入“批量服务”页面。

**图 3-119** 删除批量服务入口

**步骤2 在待删除的批量服务所在行，单击“删除”。**

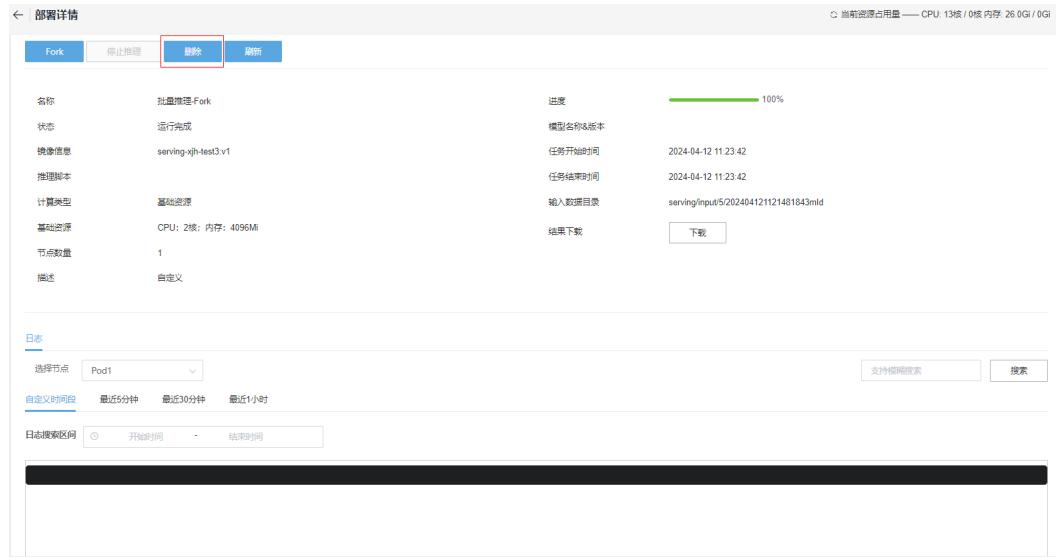
弹出提示框。

**图 3-120** 删除批量服务



## 说明

您也可以单击服务名称进入“部署详情”页面后，单击“删除”。



**步骤3** 单击“确定”，删除该服务。

### 须知

该操作将会删除该任务的所有信息且删除后无法恢复，请谨慎操作！

### ----结束

#### 3.1.7.4.7 下载批量服务结果

用户可在“批量服务”列表或部署详情页面对状态为“运行完成”的推理结果进行下载。

### 操作步骤

**步骤1** 选择“推理服务 > 批量服务”。

进入“批量服务”页面。

**图 3-121** 批量服务结果下载入口

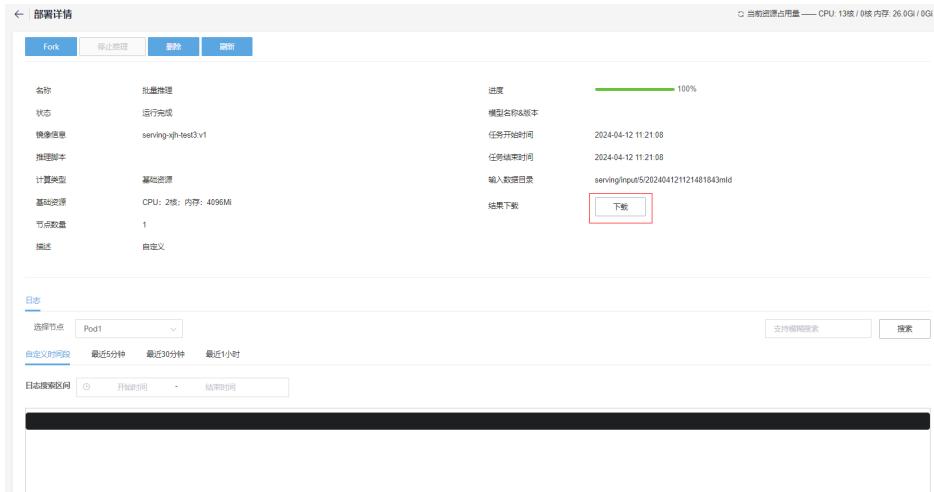


**步骤2** 在待操作的批量服务所在行，单击“结果下载”。

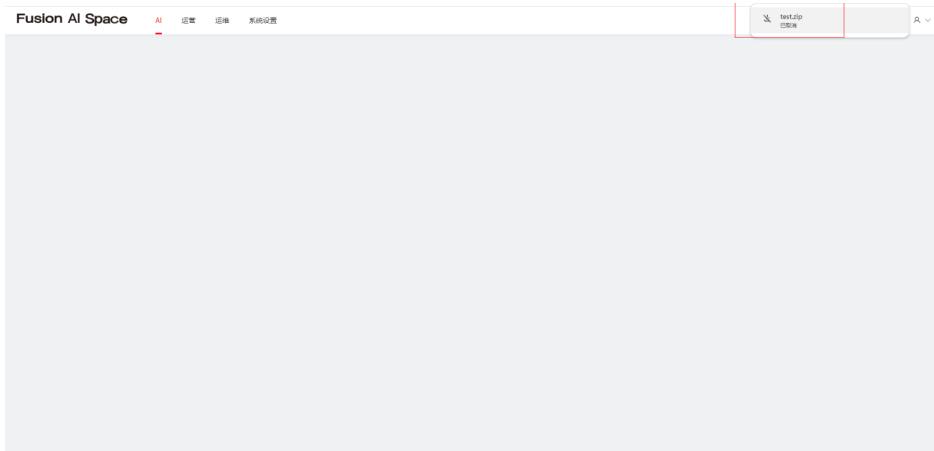
将文件下载到本地。若文件名称中包含中文字符，则下载至本地打开可能存在报错。

## 📖 说明

- 您也可以通过单击服务名称进入“部署详情”页面，单击“下载”，下载文件到本地。



- 下载文件时，如下图所示，可能会遭遇会话超时的问题，从而导致下载过程中断失败。为了解决这一问题，可以由超级管理员账户登录系统，在“系统设置 > 安全策略”界面中延长“会话超时时间”，设置完成后重新登录以下载文件。具体操作步骤请参见《AI Space 管理员指南》中的“安全策略”章节。



----结束

## 3.1.8 镜像管理

镜像是算法的载体。AI Space支持用户上传或在线制作用于训练、数据标注、可视化等操作的镜像，镜像上传至AI Space后，用户可以对镜像进行下载、分享。

### 3.1.8.1 查看镜像

#### 操作步骤

**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 选择“我的镜像”页签，查看我的镜像。

在“我的镜像”界面中可以查看到ID、创建方式、架构类型、镜像目录、镜像名称、镜像版本号、镜像地址、镜像用途、镜像描述、镜像状态、上传时间等信息。

图 3-122 我的镜像



### 步骤3 选择“公共镜像”页签，查看公共镜像。

在“公共镜像”界面中可以查看到创建方式、架构类型、镜像目录、镜像名称、镜像版本号、镜像地址、镜像用途、镜像描述、镜像状态、上传时间等信息。

图 3-123 公共镜像



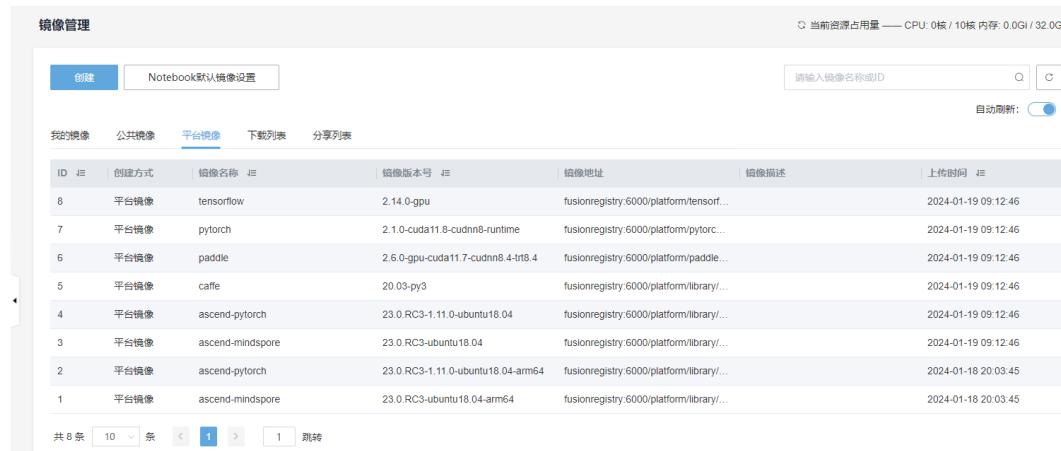
### 说明

- 组织管理员可对公共镜像进行创建，编辑，下载，删除操作；部门管理员可对公共镜像进行编辑，下载操作；普通管理员可对公共镜像进行下载操作。
- 在搜索框中输入镜像名或镜像ID，可对镜像进行搜索。

### 步骤4 选择“平台镜像”页签，查看平台镜像。

“平台镜像”界面展示了安装时预置的镜像，包括Pytorch、Tensorflow等多种常用AI开发框架镜像，用户可以直接使用平台镜像创建对应的任务或服务，方便开发者快速开展AI业务。用户可以在该界面查看平台镜像的ID、创建方式、镜像名称、镜像版本号、镜像地址、镜像描述和上传时间等信息。

图 3-124 平台镜像



----结束

### 3.1.8.2 创建镜像

AI Space提供了三种镜像创建方式，以适应不同的使用场景和需求：

- **录入镜像**：在AI Space上录入已上传至镜像仓库中的镜像信息。支持外部镜像和内部镜像。  
    外部镜像：指外部（第三方）镜像仓库中的镜像。用户可以将这些外部仓库中的镜像导入到AI Space中。  
    内部镜像：指用户自行准备并配置好的镜像，在本地加载后推送（push）到AI Space内部镜像仓库中。
- **上传镜像**：用户可以直接从本地上传准备好的镜像压缩文件，或者通过文件管理功能导入已有的镜像文件。
- **在线制作镜像**：用户可以上传或在线编辑Dockerfile文件，并上传必要的镜像依赖文件。

#### 说明

- 上传镜像和在线制作镜像将占用组织镜像额度，组织内所有用户共享该额度。超级管理员用户可在组织管理界面配置镜像额度。
- 无论是录入镜像（后台import镜像包）还是上传镜像，在镜像导入过程中，如果镜像文件较大，可能会导致管理节点的IO资源被占满。因此，建议在执行此操作时，不进行业务操作。

#### 3.1.8.2.1 录入镜像

### 加载镜像并 push 到镜像仓库

若为内部镜像，录入镜像前请参考以下步骤将镜像push到镜像仓库。

**步骤1** 在任意节点执行以下命令，加载镜像。

```
ctr -n=k8s.io images import XXX.tar
```

#### 说明

XXX表示镜像包名称。

**步骤2** 为镜像打上tag。

```
ctr -n k8s.io images tag docker.io/library/yolov5_images:v1 fusionregistry:6000/{组织code}/yolo:yolov5;
```

### 📖 说明

- docker.io/library/yolov5\_images:v1为本地镜像名称:tag，{组织code}/yolo/yolo:v1为组织code/自定义的目录/自定义的镜像名:tag，用户请根据实际情况进行修改。
- 组织code即用户所属组织的组织标识。使用超级管理员/组织管理用户登录系统后，可依次单击“系统设置 > 用户 > 组织管理/用户管理”查看对应的组织标识。

### 步骤3 push镜像到镜像仓库。

```
ctr -n k8s.io image push fusionregistry:6000/{组织code}/yolo/yolo:v1 --skip-verify --user  
ArtifactRegistryService:GZe8vNSpDTRDdUV6;
```

### 📖 说明

{组织code}/yolo/yolo:v1为组织code/自定义的目录/自定义的镜像名:tag，用户请根据实际情况进行修改。

### 步骤4 执行以下命令，查询镜像文件。

```
crtctl images
```

查询出的结果为加载的镜像地址以及镜像tag地址，录入镜像信息时填写的镜像地址为tag地址。

```
[root@master install]# crtctl images | grep yolov5  
docker.io/library/yolov5_images v1 7864cf8c9f3c8 27GB  
[root@master install]#
```

----结束

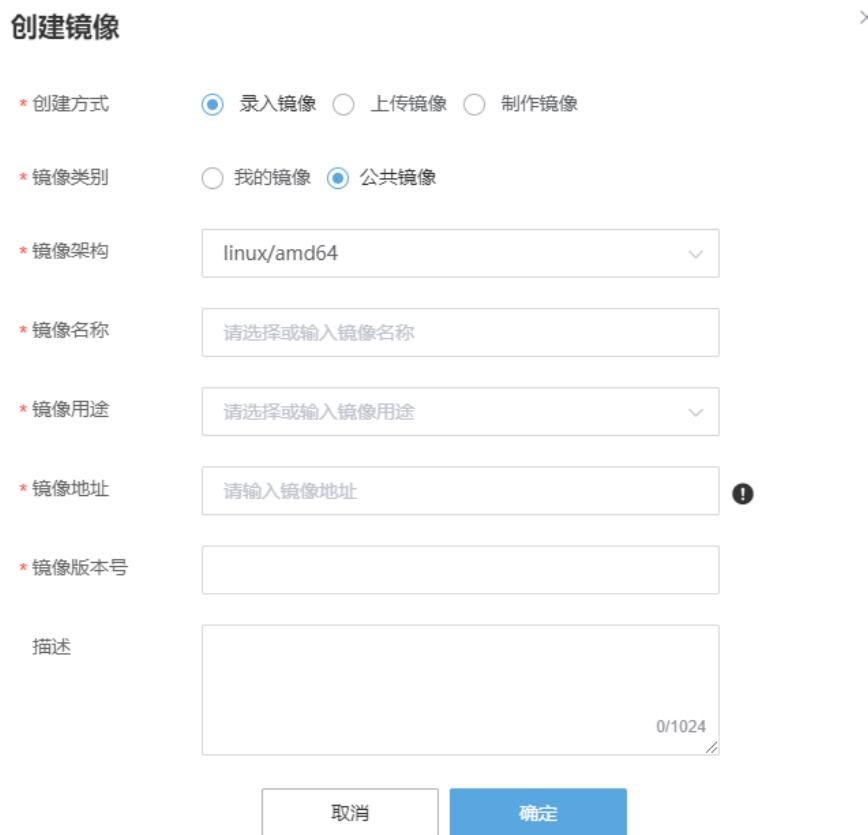
## 操作步骤

### 步骤1 选择“镜像管理”。

进入“镜像管理”界面。

### 步骤2 单击“创建”。

弹出“创建镜像”提示框。

**图 3-125 创建镜像**

**步骤3 填写镜像参数信息。**

**表 3-35 参数说明**

| 参数   | 说明  |
|------|---|
| 创建方式 | 选择创建方式为录入镜像。  |
| 镜像类别 | 选择镜像的类别，包括： <ul style="list-style-type: none"><li>我的镜像</li><li>公共镜像</li></ul>               |
| 镜像架构 | 选择镜像的架构，包括： <ul style="list-style-type: none"><li>linux/amd64</li><li>linux/arm64</li></ul> |
| 镜像名称 | 输入镜像名称，支持英文、数字、下划线和英文横杠。  |

| 参数    | 说明  |
|-------|---|
| 镜像用途  | 选择或输入镜像的用途。 <ul style="list-style-type: none"><li>● Notebook</li><li>● 训练任务</li><li>● 推理服务</li><li>● 标注服务</li><li>● 可视化</li><li>● 大模型</li></ul>   |
| 镜像地址  | 输入镜像地址。 <ul style="list-style-type: none"><li>● 内部镜像地址标准格式：fusionregistry:6000/组织code/命名空间/镜像名:镜像版本号。<br/>内部镜像地址示例：fusionregistry:6000/xfusion/enlin/notebook:v1</li><li>● 外部镜像地址标准格式：镜像仓库域名/命名空间/镜像名:镜像版本号。<br/>外部镜像地址示例：registry.cn-hangzhou.aliyuncs.com/enlin/notebook:v1</li></ul> |
| 镜像版本号 | 输入镜像版本号，镜像版本号仅支持字母、数字、英文横杠、英文句号和下划线。  |
| 描述    | 输入镜像的描述。  |

**步骤4** 单击“确定”。

页面提示创建成功。

----结束

### 3.1.8.2.2 上传镜像

#### 操作步骤

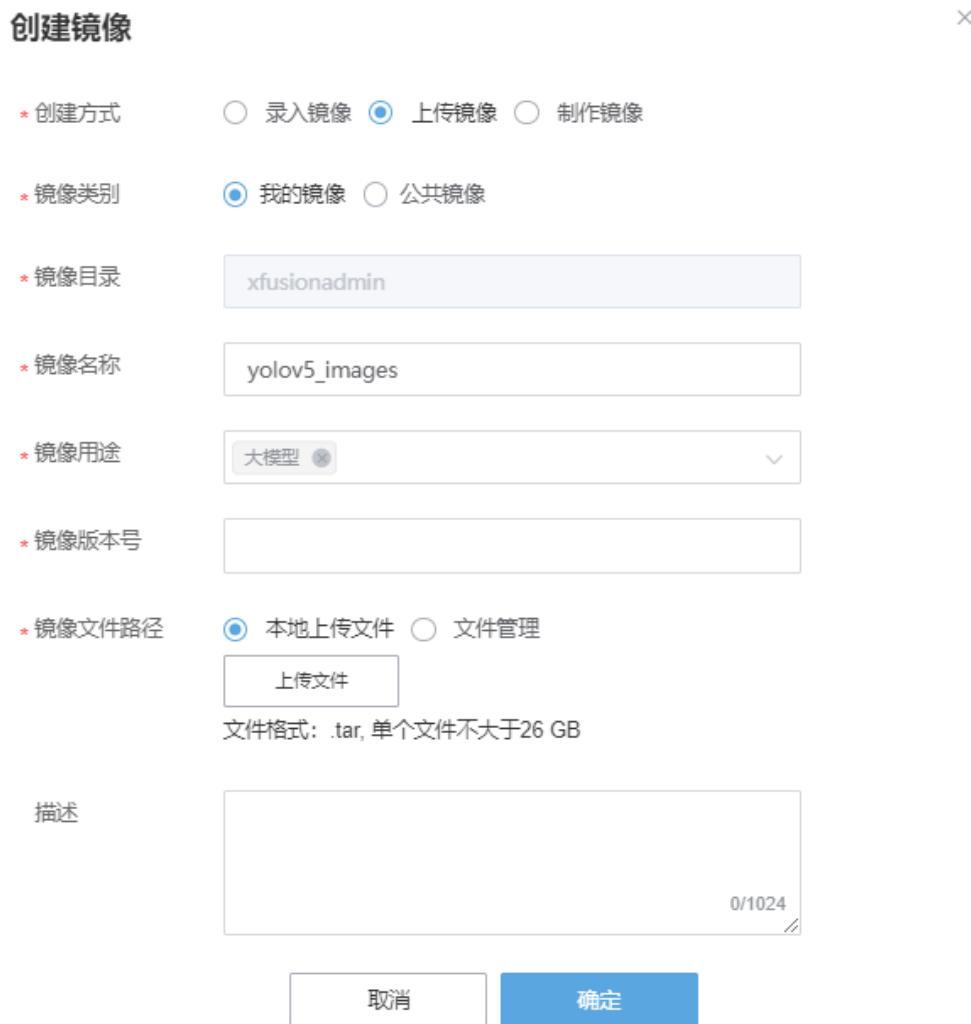
**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 单击“创建”。

弹出“创建镜像”提示框。

图 3-126 创建镜像



步骤3 填写镜像参数信息。

表 3-36 参数说明

| 参数   | 说明   |
|------|--|
| 创建方式 | 选择创建方式为上传镜像。   |
| 镜像类别 | 选择镜像的类别，包括： <ul style="list-style-type: none"><li>我的镜像</li><li>公共镜像</li></ul>                |
| 镜像目录 | 镜像存储的目录。<br>注意：当首次通过“上传镜像”或“制作镜像”创建镜像时，需输入一个镜像存放的目录，在此之后通过“上传镜像”创建镜像，系统将自动填充先前输入的目录路径，且不可修改。 |
| 镜像名称 | 输入镜像名称，支持英文、数字、下划线和英文横杠。   |

| 参数     | 说明  |
|--------|---|
| 镜像用途   | 选择或输入镜像的用途。 <ul style="list-style-type: none"><li>● Notebook</li><li>● 训练任务</li><li>● 推理服务</li><li>● 标注服务</li><li>● 可视化</li><li>● 大模型</li></ul>                                 |
| 镜像版本号  | 输入镜像版本号，镜像版本号仅支持字母、数字、英文横杠、英文句号和下划线。  |
| 镜像文件路径 | 支持从本地上传文件和从文件管理上传文件两种方式。 <ul style="list-style-type: none"><li>● 选择“本地上传文件”，单击“上传文件”上传.tar格式压缩镜像文件，大小要求不超过26GB。</li><li>● 选择“文件管理”，单击“文件管理”弹出弹窗。选择.tar格式镜像或目录，单击“确定”。</li></ul> |
| 描述     | 输入镜像的描述。  |

**步骤4** 单击“确定”。

页面提示创建成功。

----结束

### 3.1.8.2.3 制作镜像

AI Space支持基于Dockerfile的自定义制作镜像方式，用户可以选择将代码和工具上传，也可以直接在线编辑。

#### □□ 说明

制作镜像功能需要镜像架构与管理节点架构统一：

- 全X86部署场景仅支持制作X86镜像。
- 全ARM部署场景仅支持制作ARM镜像。
- X86（管理节点）、ARM异构部署场景下仅支持制作X86镜像
- ARM（管理节点）、X86异构部署场景下仅支持制作ARM镜像。

## 操作步骤

**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 单击“创建”。

弹出“创建镜像”提示框。

图 3-127 创建镜像



步骤3 填写镜像参数信息。

表 3-37 参数说明

| 参数   | 说明   |
|------|--|
| 创建方式 | 选择创建方式为制作镜像。   |
| 镜像类别 | 选择镜像的类别，包括： <ul style="list-style-type: none"><li>我的镜像</li><li>公共镜像</li></ul>                |
| 镜像目录 | 镜像存储的目录。<br>注意：当首次通过“上传镜像”或“制作镜像”创建镜像时，需输入一个镜像存放的目录，在此之后通过“制作镜像”创建镜像，系统将自动填充先前输入的目录路径，且不可修改。 |

| 参数        | 说明  |
|-----------|---|
| 镜像名称      | 输入镜像名称，支持小写英文、数字、下划线和英文横杠。  |
| 镜像用途      | 选择或输入镜像的用途。 <ul style="list-style-type: none"><li>● Notebook</li><li>● 训练任务</li><li>● 推理服务</li><li>● 标注服务</li><li>● 可视化</li><li>● 大模型</li></ul>   |
| 镜像版本号     | 输入镜像版本号，镜像版本号仅支持字母、数字、英文横杠、英文句号和下划线。  |
| 上传文件/编辑文件 | 用户可以选择上传本地Dockerfile文件或在线编辑Dockerfile文件。 <ul style="list-style-type: none"><li>● 用户在左侧下拉框选择“上传文件”后，可单击右侧“上传文件”按钮，上传本地Dockerfile文件，单个文件不大于26GB。</li><li>● 用户在左侧下拉框选择“编辑文件”后，可在右侧编辑框内在线编辑Dockerfile文件。</li></ul> <p><b>说明</b><br/>Dockerfile是用于Docker镜像的文本文件，包含所有需要用于创建镜像的命令，例如：指定基础镜像、安装依赖的软件、配置环境变量、添加文件和目录、定义容器启动时运行的命令等。<br/>关于Dockerfile文件的编写方法，请参考<a href="#">Docker官网</a>。</p> |
| 制作上传文件    | 上传制作镜像时所需依赖文件。 <ol style="list-style-type: none"><li>1. 单击“上传文件”，弹出上传文件窗口。</li><li>2. 单击“上传文件”，选择本地文件，单击“下一步”。文件格式不限，单个文件不大于26GB。</li><li>3. 文件上传成功后，单击“完成”。</li></ol>  |
| 描述        | 输入镜像的描述。  |

**步骤4** 单击“确定”。

页面提示创建成功。

----结束

### 3.1.8.3 编辑镜像

若用户创建镜像后，需要对镜像参数进行修改，可参考以下步骤。

#### 操作步骤

**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 在待编辑的镜像所在行，单击“编辑”。

弹出“修改镜像”界面。

**步骤3** 参见**3.1.8.2 创建镜像**，对镜像参数信息进行编辑修改。

- 若创建方式为“录入镜像”，用户可以修改镜像架构、镜像名称、镜像用途、镜像地址、镜像版本号和描述。

**图 3-128 修改镜像**



- 若创建方式为“上传镜像”或“制作镜像”，用户可以修改镜像用途和镜像描述。

图 3-129 修改镜像



**步骤4** 单击“确定”。

----结束

#### 3.1.8.4 设置 Notebook 默认镜像

##### 操作步骤

**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 单击“Notebook默认镜像设置”。

弹出“Notebook默认镜像设置”提示框。

图 3-130 设置 Notebook 默认镜像



**步骤3** 选择要设置的默认镜像和镜像版本信息。

**步骤4** 单击“确定”。

设置完默认镜像，在“算法开发 > 算法管理 > 我的算法”中单击“在线编辑”，会默认使用设置的镜像进行启动。

**图 3-131 Notebook 默认镜像启动**

| ID | 名称      | 描述   | 模型类别 | 创建时间                | 操作   |
|----|---------|------|------|---------------------|--|
| 48 | ppp     |      |      | 2023-09-05 19:50:17 | <a href="#">在线编辑</a> <a href="#">创建训练任务</a> <a href="#">更多</a> |
| 47 | 001     |      |      | 2023-09-05 14:47:22 | <a href="#">在线编辑</a> <a href="#">创建训练任务</a> <a href="#">更多</a> |
| 46 | 可视化1380 |      |      | 2023-09-04 14:16:59 | <a href="#">在线编辑</a> <a href="#">创建训练任务</a> <a href="#">更多</a> |
| 44 | 算法上传-1  | 图像分类 |      | 2023-09-04 10:34:35 | <a href="#">在线编辑</a> <a href="#">创建训练任务</a> <a href="#">更多</a> |

----结束

### 3.1.8.5 下载镜像

#### 操作步骤

**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 在待下载的镜像所在行，单击操作列的“下载准备”。

**图 3-132 下载准备**

| ID | 创建方式 | 购买类型        | 镜像目录          | 镜像名称 | 镜像版本号             | 镜像地址     | 镜像用途 | 镜像描述 | 镜像状态                | 上传时间   | 操作 |
|----|------|-------------|---------------|------|-------------------|----------|------|------|---------------------|--|----|
| 2  | 导入镜像 | linux/amd64 | yolov5_images | v1   | docker://libra... | Notebook |      | 成功   | 2024-02-19 15:49:55 | <a href="#">编辑</a> <a href="#">下载准备</a> <a href="#">更多</a> |    |

**步骤3** 镜像将显示在下载列表界面，下载状态变为“准备中”。

**图 3-133 下载准备中**

| ID | 创建方式 | 购买类型        | 镜像名称 | 镜像版本号 | 镜像地址                                | 下载状态 | 操作                 |
|----|------|-------------|------|-------|-------------------------------------|------|--------------------|
| 55 | 分享镜像 | 12121231234 | v1   |       | fusionregistry:5000/test_dir/121... | 准备中  | <a href="#">操作</a> |

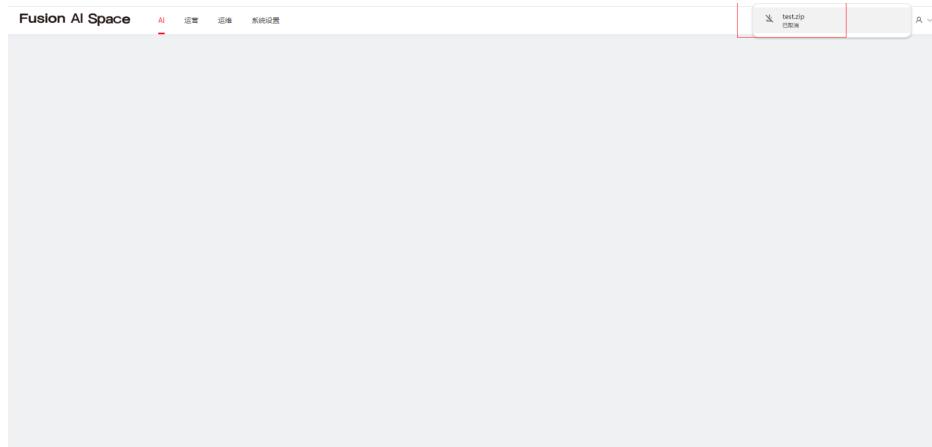
**步骤4** 下载准备完成后，状态变为“准备完成”，单击操作列“下载tar”，镜像tar包将下载至本地。

图 3-134 下载 tar



## 说明

下载文件时，如下图所示，可能会遭遇会话超时的问题，从而导致下载过程中断失败。为了解决这一问题，可以由超级管理员账户登录系统，在“系统设置 > 安全策略”界面中延长“会话超时时间”，设置完成后重新登录以下载文件。具体操作步骤请参见《AI Space 管理员指南》中的“安全策略”章节。



----结束

## 删除下载镜像文件

**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 选择“下载列表”页签。

进入下载列表界面。

**步骤3** 选择需要删除的下载文件所在行，单击操作列的“删除”。

图 3-135 删除下载镜像文件



弹出操作确认框。

**步骤4** 单击“确定”，完成删除。

----结束

### 3.1.8.6 转为公共镜像

镜像管理支持将个人镜像转为公共镜像。

**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 单击“我的镜像”页签。

进入“我的镜像”界面。

**步骤3** 在需要转公共的镜像所在行，单击操作列的“更多 > 转公共”。

图 3-136 转为公共镜像



弹出操作确认框。

**步骤4** 单击“确定”。

----结束

### 3.1.8.7 分享镜像

用户可以参考以下步骤将自己创建的镜像分享给相同组织下的其他用户。

**步骤1** 选择“镜像管理”。

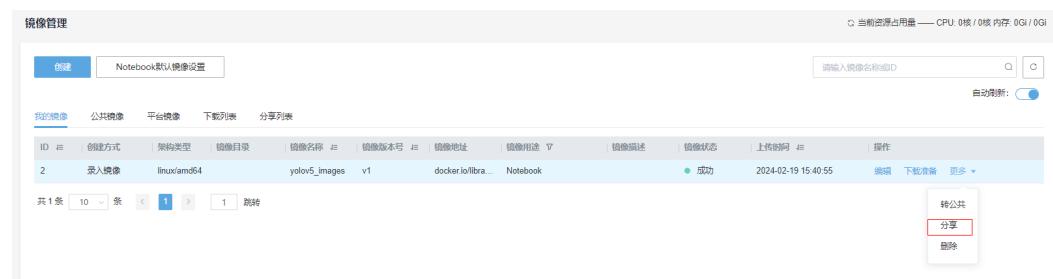
进入“镜像管理”界面。

**步骤2** 单击“我的镜像”页签。

进入“我的镜像”界面。

**步骤3** 在需要分享给其他用户的镜像所在行，单击操作列的“更多 > 分享”。

图 3-137 分享镜像



弹出“分享镜像”界面。

**步骤4** 从下拉列表中选择“被分享用户”。

**图 3-138 选择被分享用户**



**步骤5** 单击“确定”，该镜像将显示在被分享用户的“分享列表”界面。

- 被分享用户单击“同意”，镜像将会显示在“我的镜像”界面。
- 被分享用户单击“拒绝”后，该条分享信息将保留，用户也可以单击“删除”删除该条分享信息。

**图 3-139 分享列表**

| Image Management |            |                                 |  |                |                     |  |
|------------------|------------|---------------------------------|--|----------------|---------------------|--|
| Create           |            | Notebook Default Image Settings | Search for image name or ID                            |                |                     |  |
|                  |            |                                 | Automatic Refresh: <input checked="" type="checkbox"/> |                |                     |  |
| ID               | Image Name | Image Version                   | Share User ID  | Share Username | Operations          |  |
| 4                | minio_02   | V1                              | 2  | Administrator  | Agree Refuse Delete |  |
| 3                | minio_01   | V1                              | 2  | Administrator  | Agree Refuse Delete |  |

----结束

### 3.1.8.8 删除镜像

#### 操作步骤

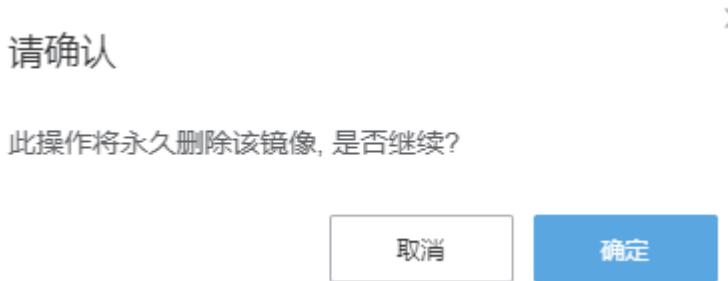
**步骤1** 选择“镜像管理”。

进入“镜像管理”界面。

**步骤2** 在待删除的镜像所在行，单击“更多 > 删除”。

弹出提示框。

图 3-140 删除镜像



**步骤3** 单击“确定”。

----结束

## 删除镜像仓库镜像

在AI Space界面执行删除镜像操作，不会删除镜像仓库的镜像文件，若用户需要删除镜像仓库的镜像文件，可参考以下步骤。

**步骤1** 在管理节点进入需要删除的镜像存放在镜像仓库的目录，示例以存放在/aaa为例。

```
cd /dfs/data/artifactregistryservice/docker/registry/v2/repositories/aaa
```

**步骤2** 执行如下命令删除镜像文件夹。

```
rm -rf <镜像名称>
```

**步骤3** 执行以下命令，查看registry仓库容器名称，图示的第一行回显结果即为registry仓库容器名称。

```
kubectl get pod -A
```

| 2- [root@master tags]# kubectl get pod -A | NAME                                   | READY | STATUS  | RESTARTS | AGE   |
|---|--|-------|---------|----------|-------|
| 2- NAMESPACE                              | ai-registry-7zjvc                      | 1/1   | Running | 0        | 3h15m |
| 2- ai-system                              | algorithm-imgprocess-7dff4dfc4f-hj6wj  | 1/1   | Running | 0        | 3h15m |
| 2- ai-system                              | algorithm-imgprocess-7dff4dfc4f-kbnnds | 1/1   | Running | 0        | 3h15m |
| 0- ai-system                              | backend-68975db8f9-8qc9q               | 1/1   | Running | 0        | 3h15m |
| 9- ai-system                              | backend-admin-56b777895d-dfgkl         | 1/1   | Running | 0        | 3h15m |
| 2- ai-system                              | backend-algorithm-9f798bdfs-fwgfv      | 1/1   | Running | 0        | 3h15m |
| 0- ai-system                              | backend-data-657c8776c9-mb92t          | 1/1   | Running | 0        | 3h15m |
| 9- ai-system                              | backend-data-task-dcd56dd94-5xxx       | 1/1   | Running | 0        | 3h15m |
| 0- ai-system                              | backend-gateway-595b76c47f-qvsdk       | 1/1   | Running | 0        | 3h15m |
| 8- ai-system                              | backend-image-86d7cc9b5b-wgkqn         | 1/1   | Running | 0        | 3h15m |
| 8- ai-system                              | backend-k8s-6474c8879b-8lsb2           | 1/1   | Running | 0        | 3h15m |
| 0- ai-system                              | backend-model-964fc9fb5-wzps5          | 1/1   | Running | 0        | 3h15m |
| 8- ai-system                              | backend-notebook-7bb77bccc-stwq9       | 1/1   | Running | 0        | 3h15m |

**步骤4** 执行以下命令进入registry仓库容器。

```
kubectl exec -it airegistry-7zjvc -n ai-system bash
```

**步骤5** 执行以下命令清理镜像。

```
/usr/local/bin/registry garbage-collect /usr/local/bin/registry-config.yml
```

**步骤6** 执行以下命令重启镜像容器。

```
kubectl delete pod airegistry-7zjvc -n ai-system
```

----结束

## 3.1.9 控制台

### 3.1.9.1 队列管理

#### 功能介绍

AI Space支持设置队列资源预留和队列容量，基于权重提供队列间资源共享。用户可以通过创建多个队列对集群的总资源进行分配管理，使用队列设置的资源处理不同的工作任务。队列管理功能需要用户部署Volcano组件，部署方法请参见《AI Space 安装指南》“安装组件”章节。

用户可以在队列管理界面查看所属组织下所有队列信息，组织管理员拥有创建、开启、关闭、编辑和删除队列的操作权限。

#### 参数说明

表 3-38 队列参数说明

| 参数             | 说明   |
|----------------|--|
| 队列名称           | 显示用户自定义的队列名称。  |
| 资源占比权重         | 显示队列的资源占比权重。<br>用户可单击资源占比权重旁的  ，对资源占比权重进行升序、降序排序。                     |
| 资源可回收          | 显示队列已使用资源量超过权重所对应的资源比例后，资源是否支持被回收。   |
| 资源预留           | 显示队列预留的资源大小。   |
| 状态             | 显示队列的状态，包括： <ul style="list-style-type: none"><li>● 可用</li><li>● 关闭</li><li>● 关闭中</li><li>● 未知</li></ul>   |
| CPU ( 已使用/容量 ) | 显示队列的CPU总容量、已使用容量和已使用占比。<br>用户可单击CPU旁的  ，按CPU使用率进行升序、降序排序。          |
| 内存 ( 已使用/容量 )  | 显示队列的内存总容量、已使用容量和已使用占比。<br>用户可单击内存旁的  ，按内存使用率进行升序、降序排序。             |
| 加速卡 ( 已使用/容量 ) | 显示队列的加速卡总容量、已使用容量和已使用占比，包括GPU卡、NPU卡。<br>用户可单击加速卡旁的  ，按使用率进行升序、降序排序。 |
| 创建时间           | 显示队列创建时间。  |

## 查看队列详细信息

**步骤1** 依次单击“控制台 > 队列管理”。

进入队列管理界面。

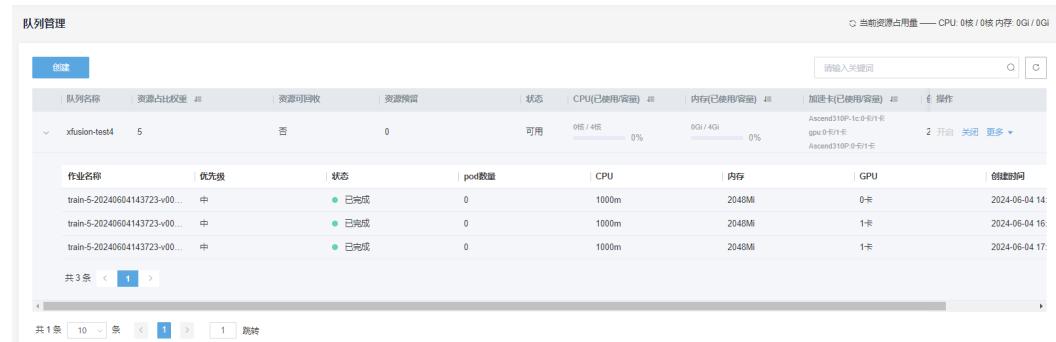
**图 3-141** 队列管理



**步骤2** 在队列管理界面可以查看队列信息，包括名称、资源占比权重、资源是否可回收、资源预留、状态、CPU使用情况、内存使用情况、GPU使用情况、NPU使用情况等信息。

**步骤3** 单击队列名称前的 ，展开节点详情，查看队列下具体的每个作业的名称，优先级，状态，pod数量，CPU、内存、GPU占用情况和创建时间。

**图 3-142** 节点详情



----结束

## 创建队列

**步骤1** 依次单击“控制台 > 队列管理”。

进入队列管理界面。

**步骤2** 单击“创建”。

弹出创建队列界面。

图 3-143 创建队列



**步骤3** 填写如下参数。

表 3-39 参数说明

| 参数     | 说明   |
|--------|--|
| 队列名称   | 用户自定义的队列名称。<br>队列名称仅支持小写字母、数字和特殊字符“-”、“_”，开头和结尾不支持“-”、“_”，并且“-”、“_”前后必须为字母或数字。 |
| 资源占比权重 | 该队列默认资源占比权重。   |
| 资源可回收  | 设置该队列资源是否可回收，默認為“否”。   |
| 容量     | 设置该队列的资源容量，包括CPU、内存、GPU、NPU。<br>容量不能超过组织可用配额。                                  |

| 参数   | 说明   |
|------|--|
| 资源预留 | 设置该队列预留的资源，包括CPU、内存、GPU、NPU。<br>资源预留值不能超过容量值。<br><b>说明</b><br>设置的资源余额值不能超过组织队列资源可预留配额。 |

**步骤4** 单击“确定”，完成队列创建。

----结束

## 编辑队列

**步骤1** 依次单击“控制台 > 队列管理”。

进入队列管理界面。

**步骤2** 在需要编辑队列的所在行，单击操作列的“更多 > 编辑”。

弹出编辑队列窗口。

**步骤3** 编辑队列参数。

**步骤4** 单击“确定”。

----结束

## 开启队列

**步骤1** 依次单击“控制台 > 队列管理”。

进入队列管理界面。

**步骤2** 在需要启用队列的所在行，单击操作列的“开启”。

弹出操作确认框。

**步骤3** 单击“确定”。

----结束

## 关闭队列

**步骤1** 依次单击“控制台 > 队列管理”。

进入队列管理界面。

**步骤2** 在需要关闭队列的所在行，单击操作列的“关闭”。

 **说明**

队列中有未完成的任务不支持关闭。

弹出操作确认框。

**步骤3** 单击“确定”。

----结束

## 删除队列

**步骤1** 依次单击“控制台 > 队列管理”。

进入队列管理界面。

**步骤2** 选择需要删除的队列，单击操作列的“更多 > 删除”。

弹出操作确认框。

### □ 说明

- 状态为可用的队列需关闭后再进行删除。
- 删除队列后，队列中的作业不会被删除。

**步骤3** 单击“确定”。

----结束

## 3.1.9.2 资源规格管理

资源规格管理为Notebook、训练任务、在线服务等提供预先设定资源，在对应的界面创建好资源规格后，创建对应的任务或服务就可以直接选择相应的资源规格。

超级管理员、组织管理员拥有资源规格管理的操作权限。

### 3.1.9.2.1 基础资源

基础资源包括CPU数量、内存大小。用户可以在该页签创建、修改、删除基础资源规格。

## 创建基础资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 选择“基础资源”页签。

**步骤3** 单击“创建”。

弹出“创建资源规格”提示框。

图 3-144 创建资源规格

The screenshot shows a dialog box titled '创建基础资源规格' (Create Basic Resource Specification). It contains several input fields:

- \* 规格名称 (Specification Name): A text input field with placeholder '请输入规格名称' (Enter specification name) and character limit '0/32'.
- \* 资源类型 (Resource Type): A dropdown menu set to '基础资源' (Basic Resources).
- \* 业务场景 (Business Scenario): A dropdown menu set to '训练任务' (Training Task).
- \* CPU数量 (CPU Number): A text input field with placeholder '请输入CPU数量' (Enter CPU number) and unit '核' (Core).
- \* 内存 (Memory): A text input field with placeholder '请输入内存大小' (Enter memory size) and unit 'Mi'. Below it, a note says '1Mi = 1024 x 1024B'.

At the bottom are two buttons: '取消' (Cancel) and '确定' (Confirm).

**步骤4** 输入规格名称、业务场景、CPU数量等信息。

#### 说明

输入的资源数量需要小于等于集群内节点上该类型卡数最大值。

**步骤5** 单击“确定”。

资源规格创建完成后，在创建Notebook、训练任务时，可以选用此规格。

----结束

## 修改基础资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 选择“基础资源”页签。

**步骤3** 在待修改信息的资源规格所在行，单击“修改”。

弹出“编辑资源规格”提示框。

图 3-145 编辑资源规格



**步骤4** 对资源规格参数信息进行编辑修改。

**步骤5** 单击“确定”。

----结束

## 删除基础资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 选择“基础资源”页签。

**步骤3** 选择需要删除的资源规格，单击“删除”。

弹出提示框。

图 3-146 删除资源规格



**步骤4** 单击“确定”。

----结束

### 3.1.9.2.2 GPU

#### ?1. 整卡

用户可以在该页签创建、修改、删除GPU整卡资源规格。

#### 创建 GPU 整卡规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“GPU > 整卡”页签。

**步骤3** 单击“创建”。

弹出“创建资源规格”提示框。

图 3-147 创建资源规格

The screenshot shows a dialog box titled '创建GPU/整卡规格' (Create GPU/Whole Card Specification). It contains several input fields:

- \* 规格名称 (Specification Name): A text input field with placeholder '请输入规格名称' (Enter specification name) and character limit '0/32'.
- \* 资源类型 (Resource Type): A dropdown menu set to 'GPU/整卡' (GPU/Whole Card).
- \* 业务场景 (Business Scenario): A dropdown menu set to '训练任务' (Training Task).
- 加速卡名称 (Accelerator Card Name): A dropdown menu placeholder '请选择加速卡名称' (Select accelerator card name).
- \* 资源名称 (Resource Name): A dropdown menu set to 'gpu'.
- \* 资源数量 (Resource Quantity): A text input field placeholder '请输入资源数量' (Enter resource quantity) followed by a suffix '-卡' (Card).

At the bottom are two buttons: '取消' (Cancel) and '确定' (Confirm).

**步骤4** 输入规格名称、业务场景、加速卡名称、资源数量等信息。

#### 说明

- 用户可以选择是否指定加速卡。
- 输入的资源数量需要小于等于集群内节点上该类型卡数最大值。

**步骤5** 单击“确定”。

资源规格创建完成后，在创建Notebook、训练任务时，可以选用此规格。

----结束

## 修改 GPU 整卡规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“GPU > 整卡”页签。

**步骤3** 在待修改信息的资源规格所在行，单击“修改”。

弹出“编辑资源规格”提示框。

图 3-148 编辑资源规格



**步骤4** 对资源规格参数信息进行编辑修改。

**步骤5** 单击“确定”。

----结束

## 删除 GPU 整卡资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“GPU > 整卡”页签。

**步骤3** 选择需要删除的资源规格，单击“删除”。

弹出提示框。

图 3-149 删 除 资 源 规 格



**步骤4** 单击“确定”。

----结束

## 2.2. MIG

用户可以在该页签创建、修改、删除MIG资源规格。

### 创建 MIG 规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“GPU > MIG”页签。

**步骤3** 单击“创建”。

弹出“创建资源规格”提示框。

图 3-150 创建资源规格



**步骤4** 输入规格名称、业务场景、资源名称、资源数量等信息。

#### □ 说明

输入的资源数量需要小于等于集群内节点上该类型卡数最大值。

**步骤5** 单击“确定”。

资源规格创建完成后，在创建Notebook、训练任务时，可以选用此规格。

----结束

## 修改 MIG 资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“GPU > MIG”页签。

**步骤3** 在待修改信息的资源规格所在行，单击“修改”。

弹出“编辑资源规格”提示框。

**图 3-151 编辑资源规格**

**步骤4** 对资源规格参数信息进行编辑修改。

**步骤5** 单击“确定”。

----结束

## 删除 MIG 资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“GPU > MIG”页签。

**步骤3** 选择需要删除的资源规格，单击“删除”。

弹出提示框。

**图 3-152 删除资源规格**

**步骤4** 单击“确定”。

----结束

### 3.1.9.2.3 NPU

#### ?1. 整卡

用户可以在该页签创建、修改、删除NPU整卡资源规格。

#### 创建 NPU 整卡规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“NPU > 整卡”页签。

**步骤3** 单击“创建”。

弹出“创建资源规格”提示框。

图 3-153 创建资源规格



**步骤4** 输入规格名称、业务场景、资源名称、资源数量等信息。

#### □ 说明

输入的资源数量需要小于等于集群内节点上该类型卡数最大值。

**步骤5** 单击“确定”。

资源规格创建完成后，在创建Notebook、训练任务时，可以选用此规格。

----结束

## 修改 NPU 整卡规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“NPU > 整卡”页签。

**步骤3** 在待修改信息的资源规格所在行，单击“修改”。

弹出“编辑资源规格”提示框。

图 3-154 编辑资源规格



**步骤4** 对资源规格参数信息进行编辑修改。

### 说明

输入的资源数量需要小于等于集群内节点上该类型卡数最大值。

**步骤5** 单击“确定”。

----结束

## 删除 NPU 整卡资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“NPU > 整卡”页签。

**步骤3** 选择需要删除的资源规格，单击“删除”。

弹出提示框。

**图 3-155** 删除资源规格

**步骤4** 单击“确定”。

----结束

## 2.2. vNPU

用户可以在该页签创建、修改、删除vNPU资源规格。

### 创建 vNPU 规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“NPU > vNPU”页签。

**步骤3** 单击“创建”。

弹出“创建资源规格”提示框。

**图 3-156** 创建资源规格A dialog box titled "创建NPU/vNPU规格" with a light gray background. It contains five input fields with validation stars: 1. "规格名称" (Specification Name) with a placeholder "请输入规格名称 0/32". 2. "资源类型" (Resource Type) with a dropdown menu showing "NPU/vNPU". 3. "业务场景" (Business Scenario) with a dropdown menu showing "训练任务". 4. "资源名称" (Resource Name) with a dropdown menu showing "请选择资源名称". 5. "资源数量" (Resource Quantity) with a placeholder "请输入资源数量". At the bottom are two buttons: a white "取消" (Cancel) button on the left and a blue "确定" (Confirm) button on the right. There is also a small "卡" icon next to the quantity input field.

**步骤4** 输入规格名称、业务场景、资源名称、资源数量等信息。

**步骤5** 单击“确定”。

资源规格创建完成后，在创建Notebook、训练任务时，可以选用此规格。

----结束

## 修改 vNPU 资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“NPU > vNPU”页签。

**步骤3** 在待修改信息的资源规格所在行，单击“修改”。

弹出“编辑资源规格”提示框。

图 3-157 编辑资源规格



**步骤4** 对资源规格参数信息进行编辑修改。

**步骤5** 单击“确定”。

----结束

## 删除 vNPU 资源规格

**步骤1** 选择“控制台 > 资源规格管理”。

进入“资源规格管理”界面。

**步骤2** 依次选择“NPU > vNPU”页签。

**步骤3** 选择需要删除的资源规格，单击“删除”。

弹出提示框。

**图 3-158** 删除资源规格



**步骤4** 单击“确定”。

----结束

## 3.2 运营

### 3.2.1 资源定价

#### 功能介绍

“资源定价”界面展示了基础资源（CPU、内存、存储），GPU资源（整卡、MIG），NPU资源（整卡、vNPU）的单价以及单价最新修改时间、操作用户。

**表 3-40** 参数说明

| 参数     | 说明   |
|--------|--|
| 资源名称   | 资源的名称。   |
| 单价     | 资源的单价。<br><b>说明</b><br>未设置的资源单价默认为0，用户创建业务任务后不会产生费用。 |
| 最新修改时间 | 资源单价的最新修改时间。   |
| 操作用户   | 修改资源单价的用户。   |

### 3.2.2 财务管理

### 3.2.2.1 收支明细

#### 功能介绍

在“收支明细”界面，组织管理员可以通过对创建时间、部门、用户、交易类型等条件筛选，查看并导出收支明细。

#### 查看收支明细

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“运营 > 财务管理 > 收支明细”。
- 进入收支明细界面。
- 步骤3** 参考[表3-41](#)选择搜索条件，单击“查询”。
- 查询收支明细。

**表 3-41 搜索条件**

| 参数   | 说明                      |
|------|-------------------------|
| 创建时间 | 交易创建的时间。                |
| 部门名称 | 部门的名称。                  |
| 交易类型 | 交易的类型，包括“全部”、“充值”和“扣费”。 |
| 操作用户 | 选择交易的操作用户。              |

----结束

#### 导出收支明细

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“运营 > 财务管理 > 收支明细”。
- 进入收支明细界面。
- 步骤3** 参考[表3-41](#)选择搜索条件，单击“查询”。
- 查询收支明细。
- 步骤4** 单击“导出”。
- 导出excel格式的收支明细。收支明细内容请参考[表3-42](#)。

**表 3-42 参数说明**

| 参数   | 说明       |
|------|----------|
| 流水号  | 交易的流水号。  |
| 交易时间 | 交易创建的时间。 |

| 参数      | 说明             |
|---------|----------------|
| 交易类型    | 交易的类型，包括充值和扣费。 |
| 交易金额(元) | 交易发生的金额。       |
| 余额(元)   | 交易完成后组织账户余额。   |
| 组织名称    | 用户所在组织名称。      |
| 部门名称    | 用户所在部门名称。      |
| 用户名称    | 用户的名称。         |

----结束

## 查看账单

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“运营 > 财务管理 > 收支明细”。  
进入收支明细界面。
- 步骤3** 参考**表3-41**选择搜索条件，单击“查询”。  
查询收支明细。
- 步骤4** 选择需要查看账单的记录，单击操作列的“查看账单”。

----结束

## 3.2.3 集群报表

### 3.2.3.1 作业详情报表

在“作业详情报表”页面，组织管理员可以查看并导出指定作业详情报表。

#### 查询作业详情报表

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“运营 > 集群报表”。  
进入集群报表界面。
- 步骤3** 选择“作业详情报表”页签。
- 步骤4** 参考**表3-43**选择搜索条件，单击“查询”。  
查询作业详情报表。

**表 3-43 搜索条件**

| 参数   | 说明                                  |
|------|-------------------------------------|
| 创建时间 | 作业创建时间。                             |
| 集群名称 | 作业所属集群名称，默认值“全部”。                   |
| 队列名称 | 作业所属队列名称，默认值“全部”。                   |
| 作业类型 | AI作业的类型，默认值“全部”。                    |
| 用户   | 用户的名称，默认值“全部”。                      |
| 作业状态 | 作业的状态，包括“等待中”、“运行中”、“暂停”、“成功”和“失败”。 |

**----结束**

## 导出作业详情报表

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运营 > 集群报表”。

进入集群报表界面。

**步骤3** 选择“作业详情报表”页签。

**步骤4** 参考**表3-43**选择搜索条件，单击“查询”。

查询作业详情报表。

**步骤5** 单击“导出”。

导出csv格式的作业详情报表。报表所含内容请参考**表3-44**。

### 📖 说明

可单击作业前  按钮，查看作业详情。

**表 3-44 参数说明**

| 参数   | 说明          |
|------|-------------|
| 作业名称 | 作业的名称。      |
| 集群   | 集群名称。       |
| 队列   | 队列名称。       |
| 用户   | 用户名称。       |
| 命名空间 | 作业Pod的名称空间。 |
| 作业类型 | 作业的类型。      |

| 参数     | 说明           |
|--------|--------------|
| 创建时间   | 作业创建时间。      |
| 开始时间   | 作业开始时间。      |
| 结束时间   | 作业结束时间。      |
| 运行时长   | 作业运行时间。      |
| 作业状态   | 作业状态。        |
| CPU核数  | 使用的CPU核数。    |
| GPU卡数  | 使用的GPU卡数。    |
| GPU卡类型 | GPU卡的类型。     |
| NPU卡数  | NPU卡数。       |
| 内存     | CPU内存，GB为单位。 |
| 显存     | GPU显存，GB单位。  |
| 控制器类型  | Pod控制器的类型。   |
| 控制器的名称 | Pod控制器的名称。   |
| Pod名称  | Pod的名称。      |
| 镜像名称   | 镜像名称。        |
| 业务类型   | 作业所属业务类型。    |
| CPU核时  | 运行时间*CPU核数。  |
| GPU卡时  | 运行时间*GPU卡数。  |

----结束

### 3.2.3.2 完成作业报表

#### 功能介绍

在“完成作业报表”页面，组织管理员可以查看并导出“完成作业统计”和“完成作业总览”相关报表。

#### 查询完成作业统计报表

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运营 > 集群报表”。

进入集群报表界面。

**步骤3** 选择“完成作业报表”页签。

**步骤4** 单击“完成作业统计”。

**步骤5** 参考**表3-45**选择搜索条件，单击“查询”。

查询完成作业统计报表。

**表 3-45** 搜索条件

| 参数   | 说明   |
|------|--|
| 采集时间 | 作业采集时间。<br><b>说明</b><br>支持通过年、月和日不同维度采集报表。                                       |
| 集群名称 | 作业所属集群名称，默认值“全部”。  |
| 队列名称 | 作业所属队列名称，默认值“全部”。  |
| 用户   | 用户的名称，默认值“全部”。   |
| 作业类型 | 作业的类型，默认值“全部”。   |
| 作业状态 | 作业的状态，包括“成功”和“失败”。   |
| 维度   | 作业报表采集维度，包括“完成时间”、“集群”、“队列”、“用户”和“作业状态”。   |
| 度量   | 作业报表采集度量，包括“已完成作业数”、“总运行时长”、“总CPU使用核数”、“总GPU使用卡数”、“总CPU使用核时”、“总GPU使用卡时”和“总等待时长”。 |

----结束

## 导出完成作业统计报表

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运营 > 集群报表”。

进入集群报表界面。

**步骤3** 选择“完成作业报表”页签。

**步骤4** 单击“完成作业统计”。

**步骤5** 参考**表3-45**选择搜索条件，单击“查询”。

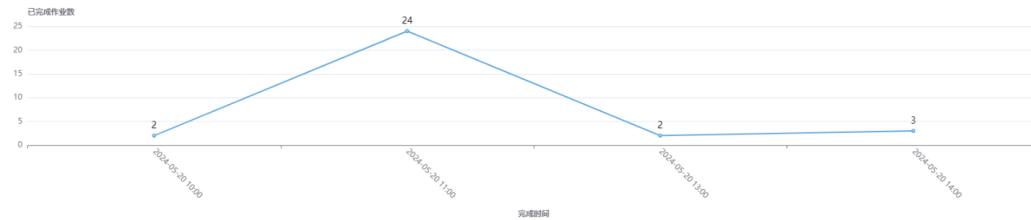
查询完成作业统计报表。

**步骤6** 单击“导出”。

导出PDF格式的完成作业统计报表。报表中可查看到设置的时间节点下完成的作业数。

## 说明书

通过具体日期查询的报表，横坐标为时间，纵坐标已完成作业数。



----结束

## 查询完成作业总览报表

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运营 > 集群报表”。

进入集群报表界面。

**步骤3** 选择“完成作业报表”页签。

**步骤4** 单击“完成作业总览”。

**步骤5** 参考**表3-46**选择搜索条件，单击“查询”。

查询完成作业总览报表。

**表 3-46** 搜索条件

| 参数   | 说明   |
|------|--|
| 采集时间 | 作业采集时间。  |
| 集群名称 | 作业所属集群名称，默认值“全部”。  |
| 队列名称 | 作业所属队列名称，默认值“全部”。  |
| 用户   | 用户的名称，默认值“全部”。   |
| 作业类型 | 作业的类型，默认值“全部”。   |
| 作业状态 | 作业的状态，包括“成功”和“失败”。   |
| 维度   | 作业报表采集维度，包括“完成时间”、“集群”、“队列”、“用户”和“作业状态”。   |
| 度量   | 作业报表采集度量，包括“已完成作业数”、“总运行时长”、“总CPU使用核数”、“总GPU使用卡数”、“总CPU使用核时”、“总GPU使用卡时”和“总等待时长”。 |

----结束

## 导出完成作业总览报表

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运营 > 集群报表”。

进入集群报表界面。

**步骤3** 选择“完成作业报表”页签。

**步骤4** 单击“完成作业总览”。

**步骤5** 参考[表3-46](#)选择搜索条件，单击“查询”。

查询完成作业总览报表。

**步骤6** 单击“导出”。

导出csv格式的完成作业总览报表。报表所含内容请参考[表3-47](#)。

**表 3-47** 参数说明

| 参数         | 说明                   |
|------------|----------------------|
| 完成作业时间     | 作业完成时间。              |
| 集群名称       | 集群名称。                |
| 队列名称       | 队列名称。                |
| 用户名称       | 用户名称。                |
| 作业状态       | 完成作业的状态，包括“成功”和“失败”。 |
| 作业个数       | 作业的个数。               |
| 作业总执行时长(秒) | 作业总执行的时长。            |
| 总CPU核数     | 使用的CPU核数。            |
| 总GPU卡数     | 使用的GPU卡数。            |
| 总CPU核时(秒)  | 运行时间*CPU核数。          |
| 总GPU卡时(秒)  | 运行时间*GPU卡数。          |
| 总等待时长(秒)   | 作业等待时长。              |

----结束

## 3.2.4 账单管理

AI Space根据用户的作业时长进行计费，作业完成后生成对应的账单。

### 3.2.4.1 我的账单

#### 功能介绍

在“我的账单”页面，可以查看并导出当前用户的账单明细。

## 查询我的账单

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“运营 > 账单管理”。
- 进入账单管理界面。
- 步骤3** 选择“我的账单”页签。
- 步骤4** 参考[表3-48](#)选择搜索条件，单击“查询”。
- 查询我的账单。

**表 3-48 搜索条件**

| 参数   | 说明                                       |
|------|--|
| 创建时间 | 账单创建的时间。                                 |
| 作业类型 | 作业的类型。                                   |
| 资源类型 | 资源的类型，包括“基础资源”、“GPU”、“MIG”、“NPU”和“vNPU”。 |

----结束

## 导出我的账单

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“运营 > 账单管理”。
- 进入账单管理界面。
- 步骤3** 选择“我的账单”页签。
- 步骤4** 参考[表3-48](#)选择搜索条件，单击“查询”。
- 查询我的账单。
- 步骤5** 单击“导出”。

导出excel格式的账单明细。明细内容请参考[表3-49](#)。

**表 3-49 参数说明**

| 参数   | 说明       |
|------|----------|
| 序号   | 账单序号。    |
| 作业名称 | 作业的名称    |
| 组织名称 | 所属组织的名称。 |
| 部门   | 所属部门名称   |
| 用户   | 创建任务的用户。 |

| 参数         | 说明            |
|------------|---------------|
| 作业类型       | 作业的类型。        |
| 创建时间       | 任务的创建时间。      |
| 开始时间       | 任务的开始时间。      |
| 结束时间       | 任务的结束时间。      |
| 运行时长       | 任务的运行时长。      |
| 资源类型       | 任务使用资源的信息。    |
| 节点数        | 任务使用的节点数。     |
| 基础资源       | 任务使用的基础资源信息。  |
| 加速资源       | 任务使用的加速卡资源信息。 |
| 加速卡单价 (元)  | 加速卡的单价。       |
| CPU单价 (元)  | CPU的单价。       |
| 内存单价 (元)   | 内存的单价。        |
| 存储单价 (元)   | 存储的单价。        |
| 存储费用 (元)   | 存储的费用。        |
| 加速卡费用 (元)  | 加速卡的费用。       |
| CPU费用 (元)  | CPU的费用。       |
| 内存费用 (元)   | 内存的费用。        |
| 任务的总费用 (元) | 任务的总费用。       |

----结束

### 3.2.4.2 所有账单

#### 功能介绍

在“所有账单”页面，组织管理员可以查看并导出组织内所有用户的账单明细。

#### 查询所有账单

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运营 > 账单管理”。

进入账单管理界面。

**步骤3** 选择“所有账单”页签。

**步骤4** 参考**表3-50**选择搜索条件，单击“查询”。

查询所有账单。

**表 3-50 搜索条件**

| 参数   | 说明                                       |
|------|--|
| 创建时间 | 账单创建的时间。                                 |
| 部门   | 部门名称。                                    |
| 操作用户 | 作业的操作用户。                                 |
| 作业类型 | 作业的类型。                                   |
| 资源类型 | 资源的类型，包括“基础资源”、“GPU”、“MIG”、“NPU”和“vNPU”。 |

----结束

## 导出所有账单

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运营 > 账单管理”。

进入账单管理界面。

**步骤3** 选择“所有账单”页签。

**步骤4** 参考**表3-50**选择搜索条件，单击“查询”。

查询所有账单。

**步骤5** 单击“导出”。

导出excel格式的账单明细。明细内容请参考**表3-51**。

**表 3-51 参数说明**

| 参数   | 说明       |
|------|----------|
| 序号   | 账单序号。    |
| 作业名称 | 作业的名称    |
| 组织名称 | 所属组织的名称。 |
| 部门   | 所属部门名称   |
| 用户   | 创建任务的用户。 |
| 作业类型 | 作业的类型。   |
| 创建时间 | 任务的创建时间。 |
| 开始时间 | 任务的开始时间。 |
| 结束时间 | 任务的结束时间。 |
| 运行时长 | 任务的运行时长。 |

| 参数        | 说明            |
|-----------|---------------|
| 资源类型      | 任务使用资源的信息。    |
| 节点数       | 任务使用的节点数。     |
| 基础资源      | 任务使用的基础资源信息。  |
| 加速资源      | 任务使用的加速卡资源信息。 |
| 加速卡单价(元)  | 加速卡的单价。       |
| CPU单价(元)  | CPU的单价。       |
| 内存单价(元)   | 内存的单价。        |
| 存储单价(元)   | 存储的单价。        |
| 存储费用(元)   | 存储的费用。        |
| 加速卡费用(元)  | 加速卡的费用。       |
| CPU费用(元)  | CPU的费用。       |
| 内存费用(元)   | 内存的费用。        |
| 任务的总费用(元) | 任务的总费用。       |

----结束

## 3.3 运维

### 3.3.1 日志

#### 3.3.1.1 审计日志

##### 功能介绍

在“审计日志”界面中，记录了当前系统成功或失败的操作日志及安全日志。在此界面中，您可以通过设置搜索条件获得指定范围的日志信息，并可将日志信息导出到本地。

- 界面显示的日志信息仅包括操作日志和安全日志。
- 通过“导出日志”按钮可以获取到操作日志、安全日志的离线列表。
  - 操作日志：用户对AI Space执行的设置类或在界面上执行普通操作类的日志。
  - 安全日志：记录安全性相关事件。例如：用户权限变更、登录或注销等信息。

##### □□ 说明

审计日志仅超级管理员和安全审计员可见。

## 查看指定范围的审计日志信息

### □ 说明

默认情况下，此界面显示所有的操作日志及安全日志。您可以通过设置高级搜索条件，获取到指定范围的日志信息。

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运维 > 日志 > 审计日志”，进入审计日志管理界面。

**步骤3** 单击列表右上方的“高级搜索”。

展开高级搜索条件设置区域。

**步骤4** 单击各项目的可选参数，设置筛选条件。

列表展示筛选后的日志信息。

### □ 说明

如需重新设置筛选条件，可单击“重置”，将筛选条件清空后重新设置。

**步骤5**（可选）单击“站点”后的，选择目标站点，筛选指定站点或全部站点的日志信息。

**步骤6** 在下方的日志列表中，查看筛选后的日志信息。

----结束

## 导出审计日志

### □ 说明

为防止日志空间溢出导致早期日志信息丢失，请定期导出日志并保存。

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“运维 > 日志 > 审计日志”，进入审计日志管理界面。

**步骤3** 单击列表左上方的“导出日志”。

弹出导出日志对话框。

**步骤4** 单击，打开日期选择器。

**步骤5** 设置“开始日期”和“结束日期”，单击“确定”。

### □ 说明

选择的日期不能晚于当前日期并且所选时间段不能超过7天。

**步骤6** 单击“确定”。

任务执行成功后，浏览器将自动下载日志文件到本地。

----结束

## 3.4 系统设置

### 3.4.1 用户

#### 3.4.1.1 用户信息

##### 功能介绍

在“用户信息”界面中，您可以查看到系统当前登录用户的名称、登录时间及角色权限，并可以对其密码进行修改。

##### 修改密码

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 用户信息”，进入当前登录用户管理界面。

**步骤3** 单击“密码修改”。

打开密码修改界面。

**步骤4** 按照提示信息输入新密码、密码确认及旧密码后，单击“确定”。

##### □□ 说明

新密码需要满足如下要求：

- 长度为8~32个字符。
- 至少包含一个空格或者以下特殊字符：  
`~!@#\$%^&\*()\_-+=|[{}];:",<,>/?
- 至少包含以下字符中的两种：
  - 小写字母：a ~ z
  - 大写字母：A ~ Z
  - 数字：0 ~ 9
- 不能包含用户名或用户名倒写。
- 新旧密码至少在2个字符位上不同。
- 不能与前N次的历史密码相同，N的取值可以参见《AI Space 管理员指南》安全策略章节获取。
- 不能为弱密码字典中的密码。

----结束

#### 3.4.1.2 用户管理

##### 功能介绍

组织管理员根据自己的实际需要对组织部门和组织用户进行管理。

在“用户”页签，可以查看组织内所有用户信息，包括用户的名称、角色、所属部门、状态等信息，并提供创建、删除、修改和禁用用户等功能。

在“部门”页签，可以查看组织内所有部门信息，包括部门的名称、描述、部门管理员、成员数量、状态等信息，并提供组织管理者创建、删除、修改、禁用用户和成员管理的功能。

**表 3-52 组织用户参数说明**

| 参数   | 说明   |
|------|--|
| 用户名称 | 组织用户的名称。   |
| 角色   | 组织用户的角色。   |
| 所属部门 | 组织用户所属的部门。   |
| 状态   | 组织用户的状态。 <ul style="list-style-type: none"><li>● 禁用</li><li>● 启用</li></ul> |
| 创建者  | 组织用户的创建者。  |
| 创建时间 | 创建该组织用户的时间。  |

**表 3-53 组织部门参数说明**

| 参数    | 说明   |
|-------|--|
| 部门名称  | 部门的名称。   |
| 描述    | 部门的描述信息。   |
| 部门管理员 | 部门的管理员。  |
| 成员数量  | 部门的成员数量。   |
| 状态    | 部门的状态。 <ul style="list-style-type: none"><li>● 启用</li><li>● 禁用</li></ul> |
| 创建者   | 创建该部门的用户。  |
| 创建时间  | 创建该部门的时间。  |
| 修改时间  | 该部门最新一次修改时间。   |

## 创建部门管理员用户/普通用户

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

**步骤3** 选择“用户”页签。

**步骤4** 单击“创建用户”。

弹出创建用户对话框。

#### 步骤5 输入用户信息。

**表 3-54 用户参数**

| 区域   | 参数   | 说明  |
|------|------|---|
| 基本信息 | 用户名  | 用户名只能包含大小写字母或数字。<br>可输入的字符串长度请参考《Fusion AI Space 管理员指南》安全策略的“帐号最小长度限制”和“帐号最大长度限制”。<br>默认长度范围为：6~32  |
|      | 密码   | 新建用户的密码。取值要求具体参见 <a href="#">3.4.1.1 用户信息</a> 。   |
|      | 密码确认 | 再次输入设置的密码。  |
| 角色权限 | 角色   | 用户具备的权限角色。<br>普通模式下可为手动创建的用户分配的角色包括： <ul style="list-style-type: none"><li>• 部门管理员，创建部门管理员后请从下拉列表选择部门管理员关联部门。只能选择已启用部门。</li><li>• 普通用户</li><li>• 自定义角色</li></ul> 当需要为用户设置多种角色时，可单击“添加权限”增加。 |

#### 步骤6 单击“确定”。

----结束

## 修改用户

#### 步骤1 登录AI Space界面。

#### 步骤2 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

#### 步骤3 选择“用户”页签。

#### 步骤4 在目标用户所在行，单击“更多 > 修改”。

#### 步骤5 编辑用户信息。

#### 步骤6 单击“确定”。

----结束

## 删除用户

#### 步骤1 登录AI Space界面。

#### 步骤2 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

#### 步骤3 选择“用户”页签。

#### 步骤4 在目标用户所在行，单击“删除”。

弹出操作确认框。

**步骤5** 单击“确定”。

----结束

## 禁用用户

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

**步骤3** 选择“用户”页签。

**步骤4** 在目标用户所在行，单击“禁用”。

弹出操作确认框。

### 说明

禁用用户后，用户将不能登录AI Space界面，当前已登录用户将自动退出。

**步骤5** 单击“确定”。

----结束

## 启用用户

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

**步骤3** 选择“用户”页签。

**步骤4** 在目标用户所在行，单击“启用”。

弹出操作确认框。

**步骤5** 单击“确定”。

----结束

## 用户配额

组织管理员可以请参考本小节步骤，为用户配置资源容量大小，用户在进行AI业务任务时不能超过该容量限制。

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

**步骤3** 选择“用户”页签。

**步骤4** 在目标用户所在行，单击“配额”。

进入配额界面，在配额界面可以查看资源配置详情。

表 3-55 资源配额详情

| 参数   | 说明                             |
|------|--------------------------------|
| 资源类型 | 资源的类型，包括基础资源、GPU、MIG、NPU、vNPU。 |
| 资源名称 | 资源的名称。                         |
| 已用配额 | 显示当前已使用资源数量及百分比。               |
| 可用配额 | 可使用配额数量。                       |
| 组织配额 | 显示用户所属组织的配额。                   |
| 部门配额 | 显示用户所属部门的配额，仅已关联部门的用户可见。       |

**步骤5** 单击可用配额列的 。

可用配额变为可编辑状态。

**步骤6** 修改可用配额。

**步骤7** 单击操作列的 ，完成修改。

----结束

## 创建部门

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

**步骤3** 选择“部门”页签。

**步骤4** 单击“创建部门”。

**步骤5** 设置部门信息参数。

表 3-56 部门信息参数说明

| 参数   | 说明  |
|------|---|
| 基本信息 |   |
| 部门名称 | 部门的名称。<br>长度为4~32个字符，只能包含中文、字母、数字、“_”、“-”和空格，且不能全为空格。 |
| 描述   | 部门的描述信息。  |
| 启用状态 | 是否启用该部门。  |

**步骤6** 单击“确定”。

----结束

## 删除部门

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。
- 步骤3** 选择“部门”页签。
- 步骤4** 在目标部门所在行，单击“删除”。  
弹出操作确认框。
- 步骤5** 单击“确定”。

----结束

## 修改部门

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。
- 步骤3** 选择“部门”页签。
- 步骤4** 在目标部门所在行，单击“更多 > 修改”。  
弹出提示框。
- 步骤5** 编辑部门的名称和描述信息。
- 步骤6** 单击“确定”。

----结束

## 启用部门

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。
- 步骤3** 选择“部门”页签。
- 步骤4** 在目标部门所在行，单击“更多 > 启用”。  
弹出操作确认框。
- 步骤5** 单击“确定”。

----结束

## 禁用部门

- 步骤1** 登录AI Space界面。
- 步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。
- 步骤3** 选择“部门”页签。
- 步骤4** 在目标部门所在行，单击“更多 > 禁用”。  
弹出操作确认框。

### 📖 说明

禁用部门后，部门成员将不能登录AI Space界面，当前已登录用户将自动退出。

**步骤5** 单击“确定”。

----结束

## 部门配额

组织管理员创建部门后，请参考本小节步骤，为各部门配置资源容量大小，部门内用户在进行AI业务任务时不能超过该容量限制。

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

**步骤3** 选择“部门”页签。

**步骤4** 在目标部门所在行，单击“配额”。

进入配额界面，在配额界面可以查看资源配置详情。

**表 3-57 资源配额详情**

| 参数   | 说明                             |
|------|--------------------------------|
| 资源类型 | 资源的类型，包括基础资源、GPU、MIG、NPU、vNPU。 |
| 资源名称 | 资源的名称。                         |
| 已用配额 | 显示当前已使用资源数量及百分比。               |
| 可用配额 | 可使用配额数量。                       |
| 组织配额 | 显示部门所属组织的配额。                   |

**步骤5** 单击“可用配额”列的 。

可用配额变为可编辑状态。

**步骤6** 修改可用配额。

**步骤7** 单击 ，完成修改。

----结束

## 成员管理

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 用户管理”，进入用户管理界面。

**步骤3** 在目标部门所在行，单击“更多 > 成员管理”。

进入成员管理界面。

**步骤4 在成员管理界面对部门成员进行管理。**

- 添加成员：单击“添加成员”，勾选需要添加的成员，单击“确定”。
- 移除成员：在目标成员所在行，单击操作列的“移除”，单击“确定”。
- 禁用成员：在目标成员所在行，单击操作列的“禁用”，单击“确定”。
- 启用成员：在目标成员所在行，单击操作列的“启用”，单击“确定”。
- 配额管理：在目标成员所在行，单击操作列的“配额”，进入配额界面，在配额界面可以查看用户的资源配置，单击可以对用户资源配置进行修改。

----结束

### 3.4.1.3 成员管理

#### 说明

仅部门管理员登录系统时，在系统设置 > 用户菜单下可查看成员管理界面。

### 功能介绍

部门管理员根据自己的实际需要对部门成员进行管理。

在“成员管理”界面，您可以查看到所属部门、部门描述、部门当前存在的成员，并提供启用、禁用、配额等功能。

**表 3-58 参数说明**

| 参数   | 说明        |
|------|-----------|
| 用户名  | 成员的用户名。   |
| 角色   | 成员的角色。    |
| 状态   | 成员的状态。    |
| 创建者  | 创建该成员的用户。 |
| 创建时间 | 创建该成员的时间。 |

### 启用成员

**步骤1 登录AI Space界面。****步骤2 依次单击“系统设置 > 用户 > 成员管理”，进入成员管理界面。****步骤3 在目标成员所在行，单击“启用”。**

弹出启用成员窗口。

**步骤4 单击“确定”。**

----结束

## 禁用成员

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 成员管理”，进入成员管理界面。

**步骤3** 在目标成员所在行，单击“禁用”。

弹出禁用成员窗口。

### □ 说明

禁用成员后，成员将不能登录AI Space界面，当前已登录用户将自动退出。

**步骤4** 单击“确定”。

----结束

## 配额管理

部门管理员请参考本小节步骤，为用户配置资源容量大小，用户在进行AI业务任务时不能超过该容量限制。

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 成员管理”，进入成员管理界面。

**步骤3** 在目标用户所在行，单击“配额”。

进入配额界面，在配额界面可以查看用户的资源配置详情。

**表 3-59 资源配额详情**

| 参数   | 说明                             |
|------|--------------------------------|
| 资源类型 | 资源的类型，包括基础资源、GPU、MIG、NPU、vNPU。 |
| 资源名称 | 资源的名称。                         |
| 已用配额 | 显示当前已使用资源数量及百分比。               |
| 可用配额 | 可使用配额数量。                       |
| 组织配额 | 显示用户所属组织的配额。                   |
| 部门配额 | 显示部门的配额。                       |

**步骤4** 单击“可用配额”列的 。

可用配额变为可编辑状态。

**步骤5** 修改可用配额，单击 ，完成修改。

----结束

### 3.4.1.4 角色管理

角色管理可以定义不同级别角色的操作权限和创建不同的角色，限制用户对系统的操作，以保障业务系统的稳定及业务数据的安全。

角色管理分为两种模式：普通模式、三员模式。

#### 3.4.1.4.1 普通模式

普通模式下，系统当前支持的角色分为三类：

- 用户手动创建的角色，仅拥有“Administrator”角色的用户拥有创建角色权限。
- 系统自动创建的角色，不可删除和修改，详情如[表3-60](#)所示。
- 多租户管理场景下自动创建的组织角色，不可删除和修改，详情如[表3-61](#)所示。

**表 3-60 系统自动创建角色**

| 角色名称             | 角色说明  | 角色权限   |
|------------------|-------|--|
| Administrator    | 超级管理员 | 系统升级、系统配置、配置部署、升级管理、能效管理、资产管理、日志管理、监控管理、设备管理、数据中心、系统软件管理、集群管理、设备证书管理、应用中心、一键巡检、集群报表、数据管理、资源定价、财务管理、账单管理、节点组、控制台。 |
| ReadOnly         | 只读用户  | 界面只读权限。  |
| ServiceOperator  | 操作员   | 进行作业下发等操作，无法查看和操作其他用户创建的内容。  |
| OpenAPI          | -     | 该角色仅具有OpenAPI接口的调用权限，用于通过BasicAuth方式调用OpenAPI接口实现机机调用能力。该角色的用户没有WebUI登录权限。                                       |
| TenantManageRole | 组织管理员 | 部门成员管理、集群报表、概览、数据管理、算法开发、训练管理、模型管理、推理服务、资源定价、财务管理、账单管理、节点组、镜像管理、控制台（队列管理、资源规格管理）。                                |

**表 3-61 组织角色**

| 角色名称                 | 角色说明  | 角色权限  |
|----------------------|-------|---|
| DepartmentManageRole | 部门管理员 | 部门成员管理、概览、数据管理、算法开发、训练管理、模型管理、推理服务、资源定价、账单管理、节点组、镜像管理、控制台（队列管理、资源规格管理）。 |
| OrdinaryUserRole     | 普通用户  | 概览、数据管理、算法开发、训练管理、模型管理、推理服务、资源定价、账单管理、节点组、镜像管理、控制台（队列管理、资源规格管理）。        |

## 创建角色

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 角色管理”，进入角色管理界面。

**步骤3** 选择“系统角色”页签。

**步骤4** 单击页面左上角的“创建角色”。

打开“创建角色”界面。

**步骤5** 设置角色的基本信息。

**表 3-62 角色基本信息**

| 参数   | 说明  |
|------|---|
| 角色名称 | 必选参数。<br>角色名称只能包含中文、字母、数字、“_”、“-”和空格，且不能全为空格。<br>字符长度范围：4 ~ 32。 |
| 描述   | 可选参数。<br>角色的描述信息。   |
| 权限   | 必选参数。<br>角色拥有的权限列表，可为新创建的角色赋予所需要的权限。                            |

**步骤6** 单击“确定”。

----结束

## 修改角色

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 角色管理”，进入角色管理界面。

**步骤3** 选择“系统角色”页签。

**步骤4** 单击指定角色后的“修改”。

进入“修改角色”界面。

**步骤5** 修改角色基本信息。

**步骤6** 单击“确定”。

完成角色修改。

----结束

## 删除角色

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 角色管理”，进入角色管理界面。

**步骤3** 选择“系统角色”页签。

**步骤4** 单击指定角色后的“删除”。

弹出操作确认对话框。

**步骤5** 单击“确定”。

----结束

### 3.4.1.4.2 三员模式

三员指的是系统管理员、安全管理员和安全审计员。三员有明确分工，三员权限互斥控制，实现管理权限的相互独立和制约。三员各自的权限对应如下：

**表 3-63** 三员各自对应权限

| 角色名称       | 角色权限   |
|------------|--|
| SysAdmin   | <ul style="list-style-type: none"><li>系统配置。</li><li>创建、删除、解锁、锁定用户(不能授予权限和激活，不能增加、删除、修改角色)。</li></ul>                                       |
| SecAdmin   | <ul style="list-style-type: none"><li>角色管理能力，可以给用户关联角色赋权，激活用户。</li><li>密码策略的制定，例如过期时间、密码长度、密码词典要求等(修改密码、密码重置等都由用户自己完成，但是要遵循此策略)。</li></ul> |
| AuditAdmin | 负责审计日志(操作日志、安全日志)的查看，导出、备份。  |

#### □□ 说明

- 系统管理员创建的用户不属于任何角色，无法登录。
- 系统管理员创建的用户需要安全管理员赋予角色后方可登录。

系统支持多种角色，为用户分配不同的权限，系统当前支持的角色分为三类：

- 用户手动创建的角色，仅安全管理员拥有创建角色权限。
- 系统自动创建的角色，不可删除和修改，详情如[表3-64](#)所示。
- 多租户管理场景下自动创建的组织角色，不可删除和修改，详情如[表3-65](#)所示。

表 3-64 系统自动创建角色

| 角色名称             | 角色说明  | 角色权限  |
|------------------|-------|---|
| SysAdmin         | 系统管理员 | 系统升级、系统配置、配置部署、升级管理、能效管理、资产管理、日志管理、作业管理、监控管理、设备管理、数据中心、系统软件管理、集群管理、设备证书管理、应用中心、数据管理、算法开发、训练管理、模型管理、镜像管理、推理服务、控制台。 |
| SecAdmin         | 安全管理员 | 角色管理权限、安全策略修改权限、用户管理权限。   |
| AuditAdmin       | 安全审计员 | 日志管理。   |
| ReadOnly         | 只读用户  | 界面只读权限。   |
| OpenAPI          | -     | 该角色仅具有OpenAPI接口的调用权限，用于通过BasicAuth方式调用OpenAPI接口实现机机调用能力。该角色的用户没有WebUI登录权限。  |
| TenantManageRole | 组织管理员 | 部门成员管理、集群报表、概览、数据管理、算法开发、训练管理、模型管理、推理服务、资源定价、财务管理、账单管理、节点组、镜像管理、控制台（队列管理、资源规格管理）。                                 |

表 3-65 组织角色

| 角色名称                 | 角色说明  | 角色权限  |
|----------------------|-------|---|
| DepartmentManageRole | 部门管理员 | 部门成员管理、概览、数据管理、算法开发、训练管理、模型管理、推理服务、资源定价、账单管理、节点组、镜像管理、控制台（队列管理、资源规格管理）。 |
| OrdinaryUserRole     | 普通用户  | 概览、数据管理、算法开发、训练管理、模型管理、推理服务、资源定价、账单管理、节点组、镜像管理、控制台（队列管理、资源规格管理）。        |

## 创建角色

- 步骤1 登录AI Space界面。
- 步骤2 依次单击“系统设置 > 用户 > 角色管理”，进入角色管理界面。
- 步骤3 选择“系统角色”页签。
- 步骤4 单击页面左上角的“创建角色”。

打开“创建角色”界面。

**步骤5** 设置角色的基本信息。

**表 3-66 角色基本信息**

| 参数   | 说明  |
|------|---|
| 角色名称 | 必选参数。<br>角色名称只能包含中文、字母、数字、“_”、“-”和空格，且不能全为空格。<br>字符长度范围：4 ~ 32。 |
| 描述   | 可选参数。<br>角色的描述信息。<br>字符长度范围：0 ~ 256。                            |
| 权限   | 必选参数。<br>角色拥有的权限列表，可为新创建的角色赋予所需要的权限。                            |

**步骤6** 单击“确定”。

----结束

## 修改角色

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 角色管理”，进入角色管理界面。

**步骤3** 选择“系统角色”页签。

**步骤4** 单击指定角色后的“修改”。

进入“修改角色”界面。

**步骤5** 修改角色基本信息。

**步骤6** 单击“确定”。

完成角色修改。

----结束

## 删除角色

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 用户 > 角色管理”，进入角色管理界面。

**步骤3** 选择“系统角色”页签。

**步骤4** 单击指定角色后的“删除”。

弹出操作确认对话框。

**步骤5** 单击“确定”。

----结束

## 3.4.2 安全

### 3.4.2.1 双因素认证

#### 功能说明

双因素认证功能是一种出于安全考虑，在登录时需要二次认证的功能。启用后，用户在登录系统的Web界面时，除了需要输入对应的密码外，还需要输入验证码，增强了系统的安全性。验证码由系统随机生成，发送到指定的邮箱、手机或企业微信中。

在“双因素认证”界面中，显示双因素认证功能的启用状态，以及使用的邮箱、短信和企业微信信息，并提供设置接口。

#### 说明

- 双因素认证功能的开启，需要SMTP功能支持。在配置之前，请参考《AI Space 管理员指南》SMTP配置章节，确认SMTP功能是使能状态。
- 配置短信或企业微信通知前，请参考《AI Space 管理员指南》短信配置和企业微信配置章节，确保已开启并配置短信或企业微信。
- “NIS用户”和“LDAP用户”不能开启该功能。

#### 操作步骤

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统设置 > 安全 > 双因素认证”，进入双因素认证管理界面。

**步骤3** 单击“修改”。

打开“配置”窗口。

**步骤4** 设置相关参数。

表 3-67 双因素认证参数说明

| 参数        | 说明  |
|-----------|---|
| 是否启用双因素认证 | <ul style="list-style-type: none"><li>• 单击  使其变为 ，表示开启双因素认证。</li><li>• 单击  使其变为 ，表示关闭双因素认证。</li></ul> |
| 认证方式      | 双因素认证方式，必选其一。 <ul style="list-style-type: none"><li>• 邮箱</li><li>• 阿里云短信</li><li>• 企业微信</li></ul>   |

| 参数     | 说明   |
|--------|--|
| 邮箱     | 接收验证码的邮箱地址，当认证方式选择为“邮箱”时配置此参数。必填。<br>邮箱地址长度不超过320个字符，不能包含空白字符以及下列特殊字符：<br>"(),.;<>[\\"   |
| 验证码    | 长度为6个字符，仅支持输入数字，必填。<br>单击“获取验证码”后，系统发送该验证码到目标邮箱、手机或企业微信，用于验证系统与邮箱的连通性。   |
| 短信模板编号 | 在阿里云创建的短信模板所对应的短信模板编号，当认证方式选择为“阿里云短信”时配置此参数。必填。<br>单击“获取短信模板内容”可查看短信模板内容，然后单击“复制文本”可一键复制短信模板内容。<br><b>须知</b><br>在阿里云创建短信模板时，需要填入此短信模板内容。 |
| 手机号    | 接收验证码的手机号，当认证方式选择为“阿里云短信”时配置此参数。必填。  |
| 企业微信账号 | 接收验证码的企业微信账号，当认证方式选择为“企业微信”时配置此参数。必填。  |

**步骤5** 单击“确定”。

----结束

### 3.4.3 节点组

AI Space支持用户将已纳管的节点划分为不同的节点组。

在节点组界面，超级管理员用户创建节点组并关联组织后，组织内用户可以查看组织关联的节点组信息，在创建AI任务时可以选择对应的节点组。

超级管理员可以进行创建、编辑和删除节点组的操作。

#### 界面参数说明

表 3-68 参数说明

| 参数          | 说明                         |
|-------------|----------------------------|
| 组名称         | 节点组的名称。                    |
| 节点数量        | 节点组内节点数量。                  |
| CPU（已分配/总数） | 节点组内CPU使用情况，包括CPU已分配数量、总数。 |
| 内存（已分配/总数）  | 节点组内内存使用情况，包括内存已分配数量、总数。   |

| 参数          | 说明   |
|-------------|--|
| 加速卡(已分配/总数) | 节点组内加速卡使用情况，包括加速卡已分配数量、总数。   |
| 共享属性        | 节点组共享属性。 <ul style="list-style-type: none"><li>● 共享：平台内所有组织共享，</li><li>● 独享：只提供给节点组关联组织使用，其他组织不可见。</li></ul> |
| 用途          | 节点组用途，包括notebook，推理任务，训练任务，标注任务，可视化任务。   |
| 关联组织        | 节点组关联组织数量。   |

单击节点组名称前的 ，将显示节点组内所有节点的详细信息

**表 3-69 参数说明**

| 参数          | 说明                       |
|-------------|--------------------------|
| 节点名         | 节点名称。                    |
| 状态          | 节点状态。                    |
| CPU(已分配/总数) | 节点CPU使用情况，包括CPU已分配数量、总数。 |
| 内存(已分配/总数)  | 节点内存使用情况，包括内存已分配数量、总数。   |
| 加速卡(已分配/总数) | 节点加速卡使用情况，包括加速卡已分配数量、总数。 |

### 3.4.4 许可证

#### 功能介绍

在“许可证”界面中，您可以查看当前许可证信息，并进行导入许可证、注销许可证操作。

**图 3-159 许可证界面**



表 3-70 参数说明

| 分区      | 参数名称     | 参数说明  |
|---------|----------|---|
| 许可证信息   | 文件类型     | 当前正在使用的许可证的状态。 <ul style="list-style-type: none"><li>商用永久</li><li>试用期：90天。</li><li>临时License</li><li>宽限期：临时License到期、临时License/商用License注销后，系统进入60天宽限期。宽限期内可使用全部功能。宽限期到期后，转化为已过期状态。</li><li>已过期</li></ul> |
|         | ESN信息    | 唯一标识设备的字符串，用以保证将许可证授权给指定设备。   |
|         | 有效时间     | License生效时长。  |
|         | 失效码      | 导入许可证文件后，执行注销操作，会生成失效码，用于注销该失效许可证，以及申请新的许可证文件。  |
|         | SnS服务到期日 | SnS维保服务到期日。   |
| 授权和使用信息 | 授权名称     | 该特性的授权名称。   |
|         | 使用值/授权值  | 该特性的使用值和授权值。  |

## 导入许可证

### 说明

编辑邮件并发送到License@xfusion.com，邮件需按如下模板提供相关信息：

- 项目名称：根据实际项目信息填写
- 合同号：CYXXXXXXXXXX（根据实际项目信息填写）
- AI Space ESN：从集群激活license页面复制
- 局点名称：根据实际项目局点信息填写
- License用途（正式或商用）：根据项目需求填写
- 软件版本：AI Space版本（实际版本号请单击右上角的查询）
- 授权节点数：根据实际项目局点集群信息填写
- 服务授权期限：X年（根据实际项目信息填写）

License受理中心收到邮件后，处理生成许可证文件，并通过License@xfusion.com发送许可证文件，请注意查收邮件。

**步骤1** 登录AI Space界面。

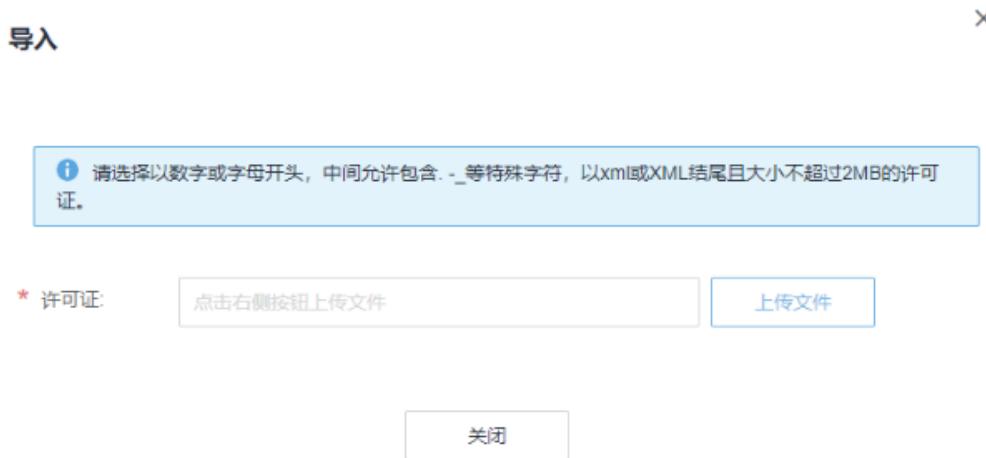
**步骤2** 依次单击“系统配置 > 许可证”，

进入“许可证”界面。

**步骤3** 单击“导入许可证”。

弹出“导入许可证”对话框。

**图 3-160 导入许可证**



**步骤4** 单击“上传文件”。

#### □ 说明

请选择以数字或字母开头，中间允许包含.-\_等特殊字符，以xml或XML结尾且压缩包大小不超过2MB的许可证。

**步骤5** 选择许可证文件。

**步骤6** 单击“导入”。

----结束

## 注销许可证

#### □ 说明

许可证注销后，从注销时刻算起，60天宽限期内该许可证功能可使用。

**步骤1** 登录AI Space界面。

**步骤2** 依次单击“系统配置 > 许可证”，

进入“许可证”界面。

**步骤3** 单击“注销许可证”。

弹出对话框。

**步骤4** 单击“确定”。

**步骤5** 记录失效码。

----结束

# 4 最佳实践

超聚变针对多种场景提供了多本最佳实践文档，为用户提供详细的操作指导，帮助用户快速开展AI业务。如有需要，请联系超聚变工程师获取。

表 4-1 最佳实践手册汇总

| 手册名称                          | 使用场景   |
|-------------------------------|--|
| LLaMA大模型微调使用指导<br>( X86+GPU ) | 介绍了在GPU ( A800 ) 上使用LLaMA-Factory进行SFT ( 有监督微调 ) 的LoRA微调验证的操作方法。 |
| Megatron-LM分布式训练框架<br>使用指导    | 介绍了使用Megatron-LM框架进行分布式训练的操作方法。                                  |

# 5 常见问题

[5.1 GPU监控为零、GPU大屏无数据](#)

[5.2 创建节点标签](#)

[5.3 进行下载操作时没有反应，下载失败](#)

[5.4 创建AI作业，提示“Read timed out executing POST http://admin/billing/created”](#)

[5.5 查看正在运行中的训练任务日志，显示“暂无日志”](#)

[5.6 获取组织的PVC Volume](#)

## 5.1 GPU 监控为零、GPU 大屏无数据

### 问题描述

用户成功安装AI Space，并纳管有GPU卡的节点后，查看GPU信息时遇到以下问题：

- 超级管理员在首页查看GPU卡数量时，显示结果为0。依次单击“监控 > 大屏 > GPU大屏”，发现GPU大屏无数据。
- 组织管理员、部门管理员和普通用户查看AI平台业务模块右上角资源监控，GPU数量显示为0。

问题可能是由于安装了不正确版本的GPU驱动所导致，可以按照以下步骤进行检查，以确认GPU驱动是否已正确安装。

### 处理步骤

**步骤1** 在有GPU卡的节点上执行nvidia-smi命令，查看回显，确认是否已正确安装GPU驱动。

Driver Version应大于等于460，小于等于525。

- 回显结果如下，表示GPU驱动版本安装正确。

```
[root@worker ~]# nvidia-smi
Tue Aug 1 17:48:39 2023
+-----+
| NVIDIA-SMI 460.106.00 | Driver Version: 460.106.00 | CUDA Version: 11.2 |
+-----+
| GPU Name Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC |
| Fan Temp Perf Pwr:Usage/Cap| Memory-Usage | GPU-Util Compute M. |
| | | | | | MIG M. |
+-----+
| 0 Tesla T4 Off 00000000:3B:00.0 Off | | | | | 0 |
| N/A 38C P8 13W / 70W | 0MiB / 15109MiB | 0% Default | N/A |
+-----+
| 1 Tesla T4 Off 00000000:86:00.0 Off | | | | | 0 |
| N/A 37C P8 13W / 70W | 0MiB / 15109MiB | 0% Default | N/A |
+-----+
+-----+
| Processes:
| GPU GI CI PID Type Process name | GPU Memory Usage |
| ID ID | |
+-----+
| No running processes found |
+-----+
```

- 回显结果如下，则表示未安装显卡驱动，可参考《AI Space 安装指南》手册中“（可选）安装NVIDIA显卡驱动”章节进行安装。

```
[root@worker1 ~]# nvidia-smi
[bash: nvidia-smi: command not found
[root@worker1 ~]# ]
```

- 回显的Driver Version不在范围内，需卸载当前NVIDIA显卡驱动后重新安装。
  - 以NVIDIA-Linux-x86\_64-525.125.06.run为例，执行如下命令卸载。  
./NVIDIA-Linux-x86\_64-525.125.06.run -uninstall
  - 执行reboot重启节点。

```
[root@worker ~]# cd /home
[root@worker ~]# ./NVIDIA-Linux-x86_64-525.125.06.run --uninstall
Uncompressing NVIDIA Accelerated Graphics Driver for Linux-x86_64 525.125.06...
.

.

.

[root@worker ~]# nvidia-smi
bash: nvidia-smi: command not found
[root@worker ~]# reboot
```

- 参考《AI Space 安装指南》手册中“（可选）安装NVIDIA显卡驱动”章节的步骤2进行安装NVIDIA显卡驱动。

## 步骤2 在管理节点下执行以下命令，重启gpu-exporter相关pod。

```
kubectl get pod | grep gpu
kubectl delete pod gpu-exporter-xx -n default
```

```
[root@master ~]# kubectl get pod | grep gpu
gpu-exporter-gl7h5 1/1 Running 0 81m
[root@master ~]# kubectl delete pod gpu-exporter-gl7h5
pod "gpu-exporter-gl7h5" deleted
[root@master ~]# kubectl get pod | grep gpu
gpu-exporter-8lsc7 1/1 Running 0 4s
```

如果重启不成功，请联系技术支持。

---结束

## 5.2 创建节点标签

### 操作步骤

**步骤1** 以root用户登录服务器。

**步骤2** 执行如下命令查询节点名。

```
kubectl get nodes
```

**步骤3** 执行如下命令给节点打标签，示例为给node167节点打上gpu-type=nvidia的标签，请根据实际情况替换。

```
kubectl label nodes node167 gpu-type=nvidia
```

**步骤4** 执行如下命令可以查看节点标签。

```
kubectl describe node 节点名
```

示例中，查询worker4节点标签为gpu=gpu。

```
[root@master ~]# kubectl describe node worker4
Name:           worker4
Roles:          worker
Labels:         beta.kubernetes.io/arch=amd64
                beta.kubernetes.io/os=linux
                gpu=gpu
                kubernetes.io/arch=amd64
                kubernetes.io/hostname=worker4
                kubernetes.io/os=linux
Annotations:   flannel.alpha.coreos.com/backend-data: {"VNI":1,"VtepMAC":"22:74:fb:fb:a8:a9"}
                flannel.alpha.coreos.com/backend-type: vxlan
                flannel.alpha.coreos.com/kube-subnet-manager: true
```

#### 说明

若用户需要删除节点标签，可参考如下命令，示例中删除worker节点的标签gpu。

```
kubectl describe nodes worker4 gpu-
```

----结束

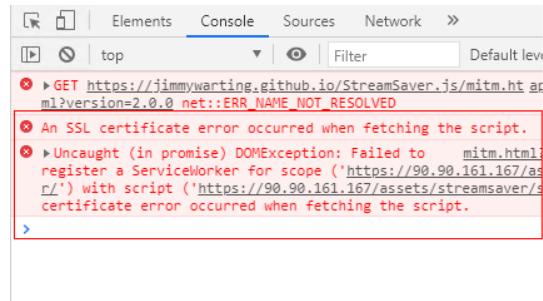
## 5.3 进行下载操作时没有反应，下载失败

### 问题描述

用户使用Google Chrome浏览器登录AI Space界面，进行以下下载操作时没有反应，下载失败。

- 导出数据集
- 下载算法
- 下载模型
- 下载镜像
- 下载批量服务结果

按F12键查看控制台，出现如图所示的报错。



可能原因为环境无法访问外部网络，浏览器阻止Service Worker注册尝试。用户可以参考解决步骤忽略证书错误。

## 解决步骤

**步骤1** 在桌面找到Google Chrome快捷方式，右键单击图标。

**步骤2** 选择“属性”标签。

**步骤3** 在“目标”文本框里可以查看程序路径。

示例：“C:\Program Files\Google\Chrome\Application\chrome.exe”。

**步骤4** 在文本后添加一个空格，并输入-ignore-certificate-errors。

修改后的示例：“C:\Program Files\Google\Chrome\Application\chrome.exe” - ignore-certificate-errors

**步骤5** 单击“确定”。

**步骤6** 重新打开Google Chrome浏览器，登录AI Space界面下载相关内容。

### 说明

若用户需要取消忽略证书错误，可删除**步骤4**中添加的内容后重启浏览器。

----结束

## 5.4 创建AI作业，提示“Read timed out executing POST http://admin/billing/created”

### 问题描述

用户在创建AI作业，如模型训练、推理服务时，出现“Read timed out executing POST http://admin/billing/created”的报错。可以参考以下步骤解决。

## 解决步骤

**步骤1** 执行以下命令进入AI Space的安装目录。

```
cd /home/AI/install/bak
```

**步骤2** 执行以下命令安装nacos客户端。

```
kubectl apply -f nacos-console.yaml
```

**步骤3** 打开浏览器，在地址栏中输入“**http://AI Space IP地址:30848/nacos**”并按“Enter”键，进入nacos界面。

**步骤4** 依次单击“配置列表 > air-prod”。

The screenshot shows the Nacos 2.0.4 configuration management interface. The left sidebar has sections like '配置管理', '历史版本', '监听管理', '服务管理', '订阅机制', '命名空间', and '集群管理'. The main area is titled '配置管理 | air-prod air-prod' and shows a table of configurations. The table has columns: Data Id, Group, 旧配置, 操作, and 指向。The Data Id column lists several YAML files: admin.yaml, ai-algorithm.yaml, ai-data-dm.yaml, ai-data-trck.yaml, ai-data.yaml, ai-modd.yaml, ai-notebook.yaml, ai-optimize.yaml, ai-pilot-cloud.yaml, and ai-serving-gateway.yaml. All entries belong to Group 'AI'. The '操作' column contains links for each file, such as '详情 | 子节点 | 编辑 | 复制 | 复原'.

**步骤5** 选择报错对应服务的yaml配置文件，单击操作列的“编辑”。

进入编辑界面。

**步骤6** 修改读取超时时间。

This screenshot shows the configuration editor for the 'ai-algorithm.yaml' file. At the top, it displays the 'Data Id: ai-algorithm.yaml' and 'Group: AI'. Below that is a '描述:' input field and a 'Beta发布:  跳过不要验证' checkbox. Under '配置格式:', there are radio buttons for TEXT, JSON, XML, YAML (which is selected), HTML, and Properties. The main content area shows the configuration code in YAML format. A red box highlights the 'connectTimeout: 5000' line, which is commented out with '#'. The code also includes 'default:' and 'readTimeout: 60000' lines. At the bottom right are '发布' (Publish) and '返回' (Return) buttons.

**步骤7** 单击“发布”。

**步骤8** 执行以下命令重启nacos。

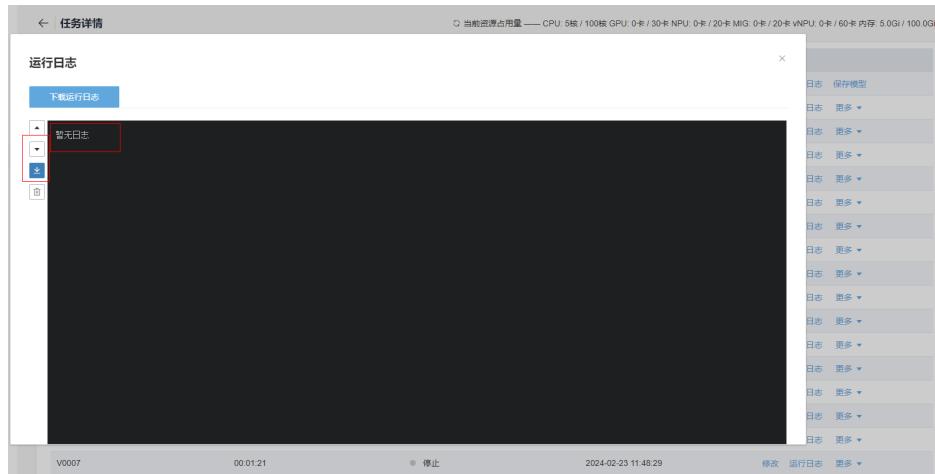
```
kubectl get pod -A | grep nacos
kubectl delete pod nacos-746cb4b5d-f8nvs -n ai-systemn
```

----结束

## 5.5 查看正在运行中的训练任务日志，显示“暂无日志”

### 问题描述

对于正在运行中的训练任务，用户在任务详情页单击“运行日志”查看任务日志。单击日志置底或自动跟随后，显示“暂无日志”。



导致问题的原因可能为训练任务日志条数超过Elasticsearch的最大返回数max\_result\_window参数，可以参考以下步骤解决。

## 解决步骤

**步骤1** 执行以下命令进入Elasticsearch容器elasticsearch-logging-0。

```
kubectl exec -it elasticsearch-logging-0 -n kube-system bash
```

**步骤2** 执行以下命令设置max\_result\_window参数。

max\_result\_window参数代表查询ES返回的条数，用户可根据需求将max\_result\_window参数设置为足够大的值。

```
curl -XPUT "http://elasticsearch-logging-0.elasticsearch-logging.kube-system.svc.cluster.local:9200/_all/_settings" -H 'Content-Type: application/json' -d '  
{  
  "index": {  
    "max_result_window": 2000000000  
  }  
}'
```

**步骤3** 参考步骤1~2，对elasticsearch-logging-1、elasticsearch-logging-2的max\_result\_window参数进行修改。

----结束

## 5.6 获取组织的 PVC Volume

PVC Volume是动态的，本小节指导用户如何获取组织的PVC Volume。

**步骤1** 使用超级管理员用户登录AI Space界面。

**步骤2** 依次单击“系统设置 > 组织管理”界面。

进入组织管理界面。

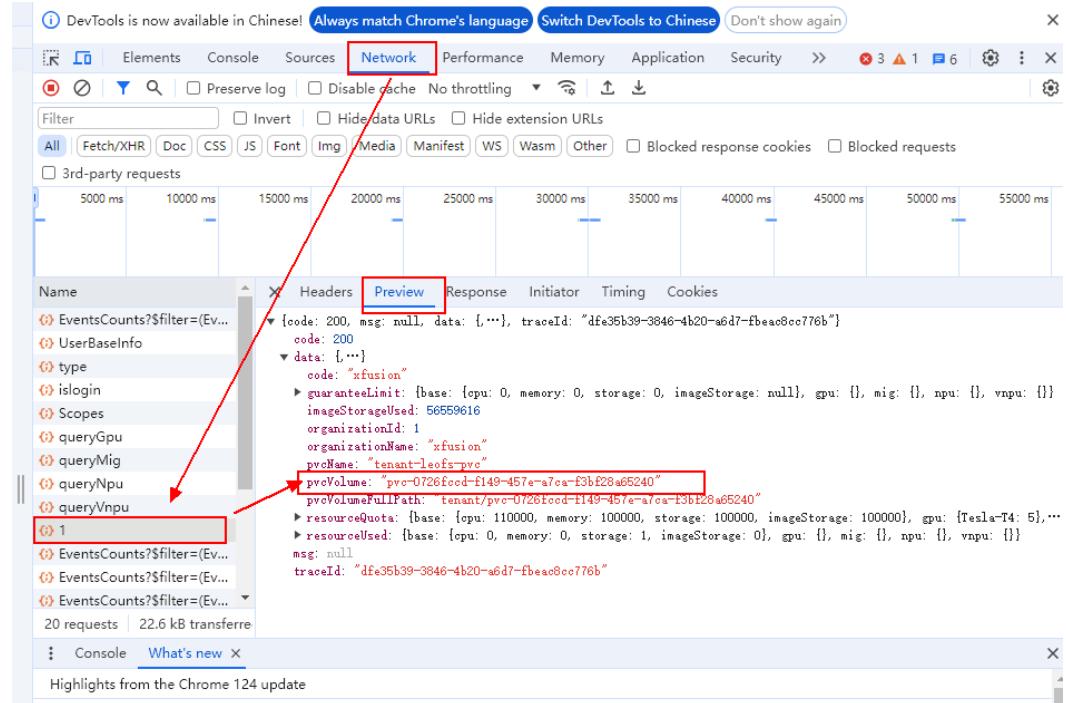
**步骤3** 右键单击界面，选择“检查”。

**步骤4** 在需要查看PVC Volume的组织所在行，单击操作列的“配额”。

**步骤5** 对组织进行配额，若已经配额过则忽略此步骤。

**步骤6** 依次选择“Network > Name > Preview”。

## 步骤7 查看PVC Volume。



----结束

# 6 术语&缩略语

## M

|      |                            |
|------|----------------------------|
| 模型训练 | 模型训练是指通过已知的数据和目标调节模型参数的过程。 |
|------|----------------------------|

## S

|      |   |
|------|---|
| 数据集  | 数据集，又称为资料集、数据集合或资料集合。在AI平台中，数据集是指包含标注的数据集合，可用于图像分类、目标检测、目标跟踪、自然语言处理等特定任务。     |
| 数据标注 | 现阶段大量的AI应用高度依赖监督式学习，通过对数据进行标注，为下游训练提供真值，常用的方法包括人工标注和机器标注配合人工编辑确认。             |
| 算法   | 算法是完成特定任务的步骤的描述，在计算机中表现为指令的有限序列。一般来说，机器学习算法可以分为监督学习、无监督学习、半监督学习、强化学习以及推荐这几大类。 |

# 7 如何获取帮助

日常维护或故障处理过程中遇到难以解决或者重大问题时，请寻求超聚变数字技术有限公司的技术支持。

## 7.1 收集必要的故障信息

在进行故障处理前，需要收集必要的故障信息。

收集的信息主要包括：

- 客户的详细名称、地址
- 联系人的姓名、电话号码
- 故障发生的具体时间
- 故障现象的详细描述
- 设备类型及软件版本
- 故障后已采取的措施和结果
- 问题的级别及希望解决的时间

## 7.2 如何使用文档

超聚变数字技术有限公司提供的指导文档能解决您在日常维护或故障处理过程中遇到的常见问题。

为了更好地解决故障，在寻求技术支持前，建议您充分使用指导文档。

## 7.3 获取技术支持

超聚变数字技术有限公司通过办事处、公司二级技术支持体系、电话技术指导、远程支持及现场技术支持等方式向用户提供及时有效的技术支持。

## 技术支持网址

查阅技术支持网站上的技术资料：访问[超聚变网站](#)。

## 案例库

参阅已有案例进行学习：[案例库](#)。

## 获取技术支持

如果在设备维护或故障处理过程中，遇到难以确定或难以解决的问题，通过文档的指导仍然不能解决，请通过如下方式获取技术支持：

- 联系超聚变数字技术有限公司客户服务中心。
  - 中国区客户可以通过以下方式联系我们：  
客户服务电话：400-009-8999  
客户服务邮箱：[support@xfusion.com](mailto:support@xfusion.com)
  - 全球各地区客户可以通过[全球服务热线](#)联系我们。
- 联系超聚变数字技术有限公司驻当地办事处的技术支持人员。