ModelArts

# 用户指南（Studio）

文档版本 01

发布日期 2024-10-11

华为云计算技术有限公司

### 商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

### 注意

您购买的产品、服务或特性等应受华为云计算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，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

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

## 华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

# 目 录



# 1 MaaS 使用场景和使用流程

ModelArts Studio大模型即服务平台（后续简称为MaaS服务），提供了简单易用的模型开发工具链，支持大模型定制开发，让模型应用与业务系统无缝衔接，降低企业AI落地的成本与难度。

当您第一次使用MaaS服务时，可以参考快速入门**使用ModelArts Studio的Llama3.1-8B模型框架实现对话问答**，了解如何在MaaS服务上的创建和部署模型。当您想更全面的了解MaaS服务的功能时，也可以参考最佳实践**在ModelArts Studio基于Llama3-8B模型实现新闻自动分类**。

**说明**

- 仅“华东二”区域支持使用ModelArts Studio大模型即服务平台（MaaS）。
- MaaS是白名单功能，如果有试用需求，请先**申请公测**。

## 应用场景

ModelArts Studio大模型即服务平台（MaaS）的应用场景：

- **业界主流开源大模型覆盖全**

  MaaS集成了业界主流开源大模型，含Llama、Baichuan、Yi、Qwen模型系列，所有的模型均基于昇腾AI云服务进行全面适配和优化，使得精度和性能显著提升。开发者无需从零开始构建模型，只需选择合适的预训练模型进行微调或直接应用，减轻模型集成的负担。

- **零代码、免配置、免调优模型开发**

  平台结合与100+客户适配、调优开源大模型的行业实践经验，沉淀了大量适配昇腾，和调优推理参数的最佳实践。通过为客户提供一键式训练、自动超参调优等能力，和高度自动化的参数配置机制，使得模型优化过程不再依赖于手动尝试，显著缩短了从模型开发到部署的周期，确保了模型在各类应用场景下的高性能表现，让客户能够更加聚焦于业务逻辑与创新应用的设计。

- **资源易获取，按需收费，按需扩缩，支撑故障快恢与断点续训**

  企业在具体使用大模型接入企业应用系统的时候，不仅要考虑模型体验情况，还需要考虑模型具体的精度效果，和实际应用成本。

  MaaS提供灵活的模型开发能力，同时基于昇腾云的算力底座能力，提供了若干保障客户商业应用的关键能力。

  保障客户系统应用大模型的成本效率，按需收费，按需扩缩的灵活成本效益资源配置方案，有效避免了资源闲置与浪费，降低了进入AI领域的门槛。

架构强调高可用性，多数据中心部署确保数据与任务备份，即使遭遇故障，也能无缝切换至备用系统，维持模型训练不中断，保护长期项目免受时间与资源损耗，确保进展与收益。

- **大模型应用开发，帮助开发者快速构建智能Agents**

  在企业中，项目级复杂任务通常需要理解任务并拆解成多个问题再进行决策，然后调用多个子系统去执行。MaaS基于多个优质昇腾云开源大模型，提供优质Prompt模板，让大模型准确理解业务意图，分解复杂任务，沉淀出丰富的多个智能Agent，帮助企业快速智能构建和部署大模型应用。

## 使用流程

**表 1-1** MaaS 使用流程

| 步骤 | 操作 | 说明 | 相关文档 |
| --- | --- | --- | --- |
| 1 | 准备工作 | 在开始使用ModelArts Studio大模型即服务平台前，需要先准备好相关依赖资源，例如创建OBS桶、创建资源池等。 | 准备MaaS资源 |
| 2 | 模型创建 | 在ModelArts Studio大模型即服务平台的“模型广场”中选择大模型模板后，需要先创建自定义大模型，才能进行模型训练和推理，才能获得更适合特定领域或任务的大语言模型。 | 在MaaS中创建模型 |
| 3 | 模型调优 | 完成数据集的准备后，可以在ModelArts Studio大模型即服务平台开始模型调优。模型调优，即使用训练数据集和验证数据集训练模型。 | 使用MaaS调优模型 |
| | 模型压缩 | 在ModelArts Studio大模型即服务平台支持对自定义模型进行模型压缩，以此提升推理服务性能、降低部署成本。 | 使用MaaS压缩模型 |
| 4 | 模型部署 | ModelArts Studio大模型即服务平台支持将自定义模型部署到计算资源上，便于在“模型体验”或其他业务环境中可以调用该模型。 | 使用MaaS部署模型服务 |
| 5 | 调用模型服务 | 在ModelArts Studio大模型即服务平台完成模型部署后，可以再其他业务环境中调用该模型服务进行预测。 | 调用MaaS部署的模型服务 |
| - | 应用体验 | ModelArts Studio大模型即服务平台提供了MaaS应用实践中心，为具体的应用场景提供一整套解决方案。 | 在MaaS应用实践中心查看应用解决方案 |

# 2 配置 MaaS 访问授权

## 2.1 配置 ModelArts 委托授权

对于所有用户（包括个人用户），需要完成ModelArts委托授权才能使用MaaS服务，否则会造成您的操作出现不可预期的错误。

- 如果您是个人用户，则不需要考虑细粒度权限问题，完成ModelArts委托授权即可使用ModelArts的所有权限。
- ModelArts平台的所有功能均通过IAM体系进行了权限管控，服务管理员可以通过标准的IAM授权动作，来对特定用户进行精细化的权限管控。

### 场景描述

MaaS服务的访问授权是通过ModelArts统一管理的，当用户已拥有ModelArts的访问授权时，无需单独配置MaaS服务的访问授权，当用户没有ModelArts的访问授权时，则需要先完成配置才能正常使用MaaS服务。

ModelArts在任务执行过程中需要访问用户的其他服务，典型的就是训练过程中，需要访问OBS读取用户的训练数据。在这个过程中，就出现了ModelArts“代表”用户去访问其他云服务的情形。从安全角度出发，ModelArts代表用户访问任何云服务之前，均需要先获得用户的授权，而这个动作就是一个“委托”的过程。用户授权ModelArts再代表自己访问特定的云服务，以完成其在ModelArts平台上执行的AI计算任务。

ModelArts提供了一键式自动授权功能，用户可以在ModelArts的权限管理功能中，快速完成委托授权，由ModelArts为用户自动创建委托并配置到ModelArts服务中。

本章节主要介绍一键式自动授权方式。一键式自动授权方式支持给IAM子用户、联邦用户（虚拟IAM用户）、委托用户和所有用户授权。

### 约束与限制

- 华为云账号
  - 只有华为云账号可以使用委托授权，可以为当前账号授权，也可以为当前账号下的所有IAM用户授权。
  - 多个IAM用户或账号，可使用同一个委托。
  - 一个账号下，最多可创建50个委托。

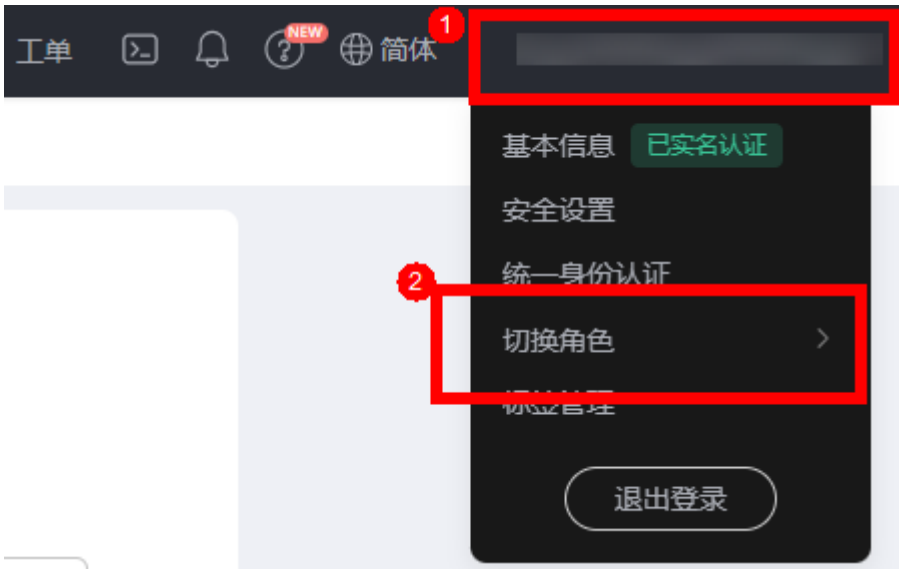
  - 对于首次使用ModelArts的新用户，请直接新增委托即可。一般用户新增普通用户权限即可满足使用要求。如果有精细化权限管理的需求，可以自定义权限按需设置。
- IAM用户
  - 如果已获得委托授权，则可以在权限管理页面中查看到已获得的委托授权信息。
  - 如果未获得委托授权，当打开“访问授权”页面时，ModelArts会提醒您当前用户未配置授权，需联系此IAM用户的管理员账号进行委托授权。

## 添加授权

1. 登录ModelArts管理控制台，在左侧导航栏选择“权限管理”，进入“权限管理”页面。
2. 单击“添加授权”，进入“访问授权”配置页面，根据参数说明进行配置。

**表 2-1** 参数说明

| 参数 | 说明 |
| --- | --- |
| “授权对象类型” | 包括IAM子用户、联邦用户、委托用户和所有用户。<br>● IAM子用户：由主账号在IAM中创建的用户，是服务的使用人员，具有独立的身份凭证（密码和访问密钥），根据账号授予的权限使用资源。IAM子用户相关介绍请参见**IAM用户介绍**。<br>● 联邦用户：又称企业虚拟用户。联邦用户相关介绍请参见**联邦身份认证**。<br>● 委托用户：IAM中创建的一个委托。IAM创建委托相关介绍请参见**创建委托**。<br>● 所有用户：该选项表示会将委托的权限授权到当前账号下的所有子账号、包括未来创建的子账号，授权范围较大，需谨慎使用。个人用户选择“所有用户”即可。 |

| 参数 | 说明 |
| --- | --- |
| “授权对象” | “授权对象类型”选择“所有用户”时不涉及此参数。<br>• IAM子用户：选择指定的IAM子用户，给指定的IAM子用户配置委托授权。<br>**图 2-1** 选择 IAM 子用户<br>  <br>• 联邦用户：输入联邦用户的用户名或用户ID。<br>**图 2-2** 选择联邦用户<br>  <br>• 委托用户：选择委托名称。使用账号A创建一个权限委托，在此处将该委托授权给账号B拥有的委托。在使用账号B登录控制台时，可以在控制台右上角的个人账号切换角色到账号A，使用账号A的委托权限。<br>**图 2-3** 委托用户切换角色<br>  |
| “委托选择” | • 已有委托：列表中如果已有委托选项，则直接选择一个可用的委托为上述选择的用户授权。单击委托名称查看该委托的权限详情。<br>• 新增委托：如果没有委托可选，可以在新增委托中创建委托权限。对于首次使用ModelArts的用户，需要新增委托。 |
| “新增委托 > 委托名称” | 系统自动创建委托名称，用户可以手动修改。 |

| 参数 | 说明 |
| --- | --- |
| “新增委托 > 授权方式” | • 角色授权：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于华为云各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。<br>• 策略授权：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。<br>角色与策略相关介绍请参考**权限基本概念**。 |
| “新增委托 > 权限配置 > 普通用户” | 普通用户包括用户使用ModelArts完成AI开发的所有必要功能权限，如数据的访问、训练任务的创建和管理等。一般用户选择此项即可。<br>可以单击“查看权限列表”，查看普通用户权限。 |
| “新增委托 > 权限配置 > 自定义” | 如用户有精细化权限管理的需求，可使用自定义模式灵活按需配置ModelArts创建的委托权限。可以根据实际需要在权限列表中勾选要配置的权限。 |

3. 然后勾选“我已经详细阅读并同意《ModelArts服务声明》”，单击“创建”，即可完成委托配置。

## 查看授权的权限列表

用户可以在“权限管理”页面的授权列表中，查看已经配置的委托授权内容。单击授权内容列的“查看权限”，可以查看该授权的权限详情。

**图 2-4** 查看权限

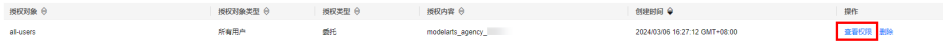
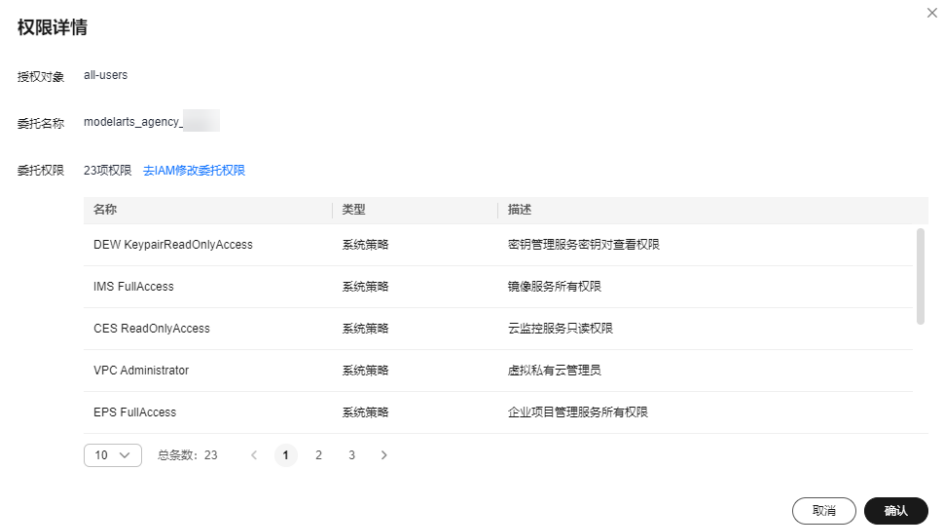



**图 2-5** 普通用户权限列表



## 修改授权的权限范围

1. 在查看授权详情时，如果想要修改授权范围，可以在权限详情页单击“去IAM修改委托权限”。

   **图 2-6** 去 IAM 修改委托权限


2. 进入IAM控制台的委托页面。找到对应的委托信息，修改该委托的基本信息，主要是持续时间。“持续时间”可以选择永久、1天，或者自定义天数，例如 30 天。

   **图 2-7** 手动创建的委托


3. 在授权记录页面单击“授权”，勾选要配置的策略，单击下一步设置最小授权范围，单击确定，完成授权修改。

   设置最小授权范围时，可以选择指定的区域，也可以选择所有区域，即不设置范围。

## 删除授权

为了更好的管理您的授权，您可以删除某一IAM用户的授权，也可批量清空所有用户的授权。

- **删除某一用户的授权**

  在“权限管理”页面，展示当前账号下为其IAM用户配置的授权列表，针对某一用户，您可以单击“操作”列的“删除”，输入“DELETE”后单击“确认”，可删除此用户的授权。删除生效后，此用户将无法继续使用ModelArts的相关功能。

- **批量清空所有授权**

  在“权限管理”页面，单击授权列表上方的“清空授权”，输入“DELETE”后单击“确认”，可删除当前账号下的所有授权。删除生效后，此账号及其所有IAM子用户将无法继续使用ModelArts的相关功能。

## 常见问题

1. 首次使用ModelArts如何配置授权?

   直接选择“新增委托”中的“普通用户”权限即可，普通用户包括用户使用ModelArts完成AI开发的所有必要功能权限，如数据的访问、训练任务的创建和管理等。一般用户选择此项即可。

2. 如何获取访问密钥AK/SK?

   如果在其他功能（例如访问模型服务等）中使用到访问密钥AK/SK认证，获取AK/SK方式请参考**如何获取访问密钥**章节。

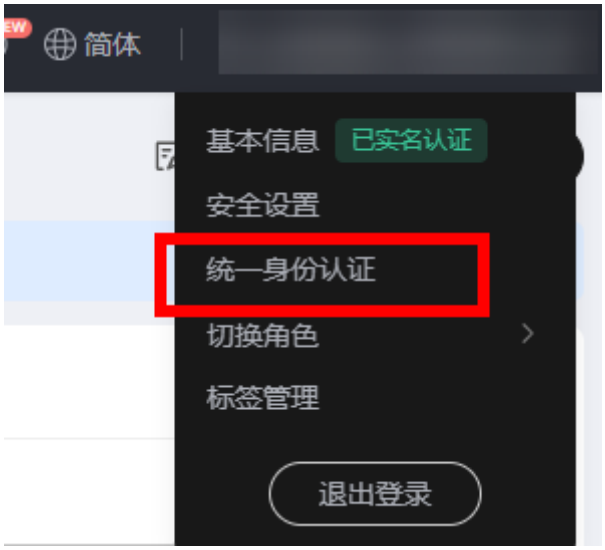
3. 如何删除已有委托列表下面的委托名称?

   **图 2-8** 已有委托

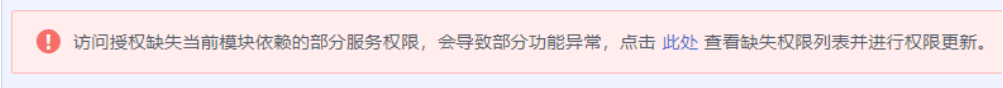
   需要前往统一身份认证服务IAM控制台的委托页面删除。

   **图 2-9** 统一身份认证

4. 进入ModelArts控制台的某个页面时，为什么会提示权限不足?

**图 2-10** 页面提示权限不足




可能原因是用户委托权限配置不足或模块能力升级，需要更新授权信息。根据界面操作提示追加授权即可。

# 2.2 创建 IAM 用户并授权使用 MaaS

**配置ModelArts委托授权**章节中介绍的一键式自动授权方式创建的委托的权限比较大，基本覆盖了依赖服务的全部权限。如果华为云账号已经能满足您的要求，则不需要创建独立的IAM用户，您可以跳过本章节，不影响您使用MaaS服务的功能。

ModelArts作为一个完备的AI开发平台，支持用户对其进行细粒度的权限配置，以达到精细化资源、权限管理之目的。这类特性在大型企业用户的使用场景下很常见。如果需要对委托授权的权限范围进行精确控制，可以参考本章节进行MaaS服务的定制化委托授权。

本章节主要介绍如何给IAM用户下的子用户配置更细粒度的权限。

## 前提条件

给用户组授权之前，请先了解用户组可以添加的使用ModelArts及其依赖服务的权限，并结合实际需求进行选择，MaaS服务支持的系统权限，请参见**表2-2**。

**表 2-2** 服务授权列表

| 待授权的服务 | 授权说明 | IAM权限设置 | 是否必选 |
| --- | --- | --- | --- |
| ModelArts | 授予子用户使用ModelArts服务的权限。<br>ModelArts CommonOperations没有任何专属资源池的创建、更新、删除权限，只有使用权限。推荐给子用户配置此权限。 | ModelArts CommonOperations | 必选 |

| 待授权的服务 | 授权说明 | IAM权限设置 | 是否必选 |
| --- | --- | --- | --- |
| | 如果需要给子用户开通专属资源池的创建、更新、删除权限，此处要勾选ModelArts FullAccess，请谨慎配置。 | ModelArts FullAccess | 可选<br>ModelArts FullAccess权限和ModelArts Common Operations权限建议二选一。 |
| OBS对象存储服务 | 授予子用户使用OBS服务的权限。ModelArts的数据管理、开发环境、训练作业、模型推理部署均需要**通过OBS进行数据中转**。 | OBS OperateAccess | 必选 |
| SWR容器镜像仓库 | 授予子用户使用SWR服务权限。ModelArts的**自定义镜像功能**依赖镜像服务SWR FullAccess权限。 | SWR OperateAccess | 必选 |
| CES云监控 | 授予子用户使用CES云监控服务的权限。通过CES云监控可以查看ModelArts的在线服务和对应模型负载运行状态的整体情况，并设置监控告警。 | CES FullAccess | 必选 |
| SMN消息服务 | 授予子用户使用SMN消息服务的权限。SMN消息通知服务配合CES监控告警功能一起使用。 | SMN FullAccess | 必选 |
| VPC虚拟私有云 | 子用户在创建ModelArts的专属资源池过程中，如果需要开启自定义网络配置，需要配置VPC权限。 | VPC FullAccess | 可选 |

## 配置 MaaS 基础操作权限

**步骤1** 创建用户组。

登录IAM管理控制台，单击“用户组>创建用户组”。在“创建用户组”界面，输入“用户组名称”单击“确定”。

**步骤2** 配置用户组权限。

在用户组列表中，单击步骤1新建的用户组右侧的“授权”，在用户组“授权”页面，您需要配置的权限如下：

1. 配置ModelArts使用权限。在搜索框搜索ModelArts。ModelArts FullAccess权限和ModelArts CommonOperations权限建议二选一。

   选择说明如下：

   - ModelArts CommonOperations没有任何专属资源池的创建、更新、删除权限，只有使用权限。推荐给子用户配置此权限。
   - 如果需要给子用户开通专属资源池的创建、更新、删除权限，此处要勾选ModelArts FullAccess，请谨慎配置。

   **图 2-11** 配置 ModelArts 使用权限

2. 配置其他依赖云服务的使用权限，此处以OBS为例，搜索OBS，勾选“OBS OperateAccess”。ModelArts训练作业中需要依赖OBS作为数据中转站，需要配置OBS的使用权限。

   更多需要配置的云服务权限请参见**表2-2**，比如SWR等，重复操作此步骤。

3. 再单击“下一步”，设置最小授权范围。单击“指定区域项目资源”，勾选待授权使用的区域，单击“确定”。
4. 提示授权成功，查看授权信息，单击“完成”。此处的授权生效需要15-30分钟。

**步骤3** 创建子用户账号。在IAM左侧菜单栏中，选择“用户”，单击右上角“创建用户”，在“创建用户”页面中，添加多个用户。请根据界面提示，填写必选参数，然后单击“下一步”。

**步骤4** 将子用户子账号加入用户组。在“加入用户组”步骤中，选择“用户组”，然后单击“创建用户”。系统将前面设置的多个用户加入用户组中。

**步骤5** **用户登录**并验证权限。

新创建的用户登录控制台，切换至授权区域，验证权限：

- 在“服务列表”中选择ModelArts，进入ModelArts主界面，选择不同类型的专属资源池，在页面单击“创建”，如果无法进行创建（当前权限仅包含ModelArts CommonOperations），表示“ModelArts CommonOperations”已生效。
- 在“服务列表”中选择除ModelArts外（假设当前策略仅包含ModelArts CommonOperations）的任一服务，如果提示权限不足，表示“ModelArts CommonOperations”已生效。
- 在“服务列表”中选择ModelArts，进入ModelArts主界面，单击“数据管理>数据集>创建数据”，如果可以成功访问对应的OBS路径，表示OBS权限已生效。
- 参考**表2-2**依次验证其他可选权限。

**----结束**

# 3 准备 MaaS 资源

在使用MaaS服务时，需要先完成OBS桶、资源池等准备工作。

## 准备 OBS 桶

在ModelArts Studio大模型即服务平台创建自定义模型、调优或压缩模型时，需要在对象存储服务OBS中创建OBS桶，用于存放模型权重文件、训练数据集或者是存放永久保存的日志。

创建OBS桶和上传文件的操作指导请参见**OBS控制台快速入门**。

**说明**

OBS桶必须和MaaS服务在同一个Region下，否则无法选择到该OBS路径。

## 准备资源池

在ModelArts Studio大模型即服务平台进行模型调优、压缩或部署时，需要选择资源池。MaaS服务支持专属资源池和公共资源池。

- 专属资源池：专属资源池不与其他用户共享，资源更可控。在使用专属资源池之前，您需要先创建一个专属资源池，然后在AI开发过程中选择此专属资源池。MaaS服务可以使用在ModelArts Standard形态下创建的专属资源池用于模型训推。创建专属资源池的操作指导请参见**创建Standard专属资源池**。

  **须知**

  MaaS服务只支持使用驱动版本是23.0.5的专属资源池，其他版本会导致任务失败。当专属资源池的驱动版本不适配时，可以参考**升级Standard专属资源池驱动**升级驱动。

- 公共资源池：公共资源池提供公共的大规模计算集群，根据用户作业参数分配使用，资源按作业隔离。MaaS服务可以使用ModelArts Standard形态下提供的公共资源池完成模型训推，按照使用量计费，方便快捷。选择公共资源池时，可以通过购买套餐包获取优惠的资源费用，请参见**购买套餐包**。

**说明**

资源池必须和MaaS服务在同一个Region下，否则无法选择到该资源池。

## 购买套餐包

MaaS服务提供了按需套餐包，用户可以提前购买按需套餐包，从而获得灵活的、更高性价比的算力资源。当购买了套餐包，在使用公共资源池运行任务时，将会优先抵扣套餐包的配额，超出当前套餐包的额度或使用时段，将自动转为按需收费。

关于套餐包的约束限制、资源包抵扣顺序和套餐包余量预警请参见**套餐包**。

购买操作如下：

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio首页单击“购买套餐包”，进入购买页面。
4. 在购买页面，选择套餐包“规格”和“购买数量”，单击“立即购买”，确认订单详情，单击“去支付”，根据界面提示完成套餐包支付。
5. 支付完成后，在ModelArts Studio大模型即服务平台创建任务时，选择套餐包规格的公共资源池，在运行任务时既可优先使用套餐包资源。

# 4 在模型广场查看模型

在模型广场页面，ModelArts Studio大模型即服务平台提供了丰富的开源大模型模板，在模型详情页可以查看模型的详细介绍，根据这些信息选择合适的模型模板，用于创建模型。

## 访问模型广场

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型广场”进入模型广场。
4. 选择模型，单击“立即使用”进入模型详情页。在模型详情页可以查看模型的详细介绍。

## 模型介绍

表4-1列举了ModelArts Studio大模型即服务平台支持的模型清单，模型详细信息请查看界面介绍。

**表 4-1** 模型广场的模型系列介绍

| 模型系列 | 模型类型 | 应用场景 | 支持语言 |
| --- | --- | --- | --- |
| GLM-4 | 文本生成 | 对话问答、长文本推理、代码生成 | 中文、英文 |
| ChatGLM3 | 文本生成 | 对话问答、数学推理、代码生成 | 中文、英文 |
| 百川 | 文本生成 | 对话问答、数学推理、代码生成、翻译 | 中文、英文 |
| Llama 2 | 文本生成 | 对话问答、智能创作、文本摘要 | 英文 |
| Llama 3 | 文本生成 | 对话问答、智能创作、文本摘要 | 英文 |
| Llama 3.1 | 文本生成 | 对话问答、智能创作、文本摘要 | 英文 |
| Yi | 文本生成 | 代码生成、数学推理、对话问答 | 中文、英文 |

| 模型系列 | 模型类型 | 应用场景 | 支持语言 |
| --- | --- | --- | --- |
| 通义千问1.5 | 文本生成 | 代码生成、数学推理、对话问答 | 英文 |
| 通义千问 | 文本生成 | 对话问答、智能创作、文本摘要、翻译、代码生成、数学推理 | 中文、英文 |
| 通义千问2 | 文本生成 | 多语言处理、数学推理、对话问答 | 英文 |

模型分为量化模型和非量化模型，其中，量化模型又包括SmoothQuant-W8A8和AWQ-W4A16两种。

- AWQ-W4A16量化模型可以由非量化模型压缩后生成，也可以直接使用开源AWQ权重。
- SmoothQuant-W8A8量化模型只能由非量化模型压缩生成。

ModelArts Studio大模型即服务平台已预置非量化模型与AWQ-W4A16量化模型的模型模板。

- 非量化模型可以支持调优、压缩、部署等操作。
- 量化模型仅支持部署操作。当需要获取SmoothQuant-W8A8量化模型时，则可以通过对非量化模型进行模型压缩获取。

# 5 在 MaaS 中创建模型

在模型广场选择模型后，需要使用模型创建一个自定义模型，才能进行模型训练、推理。

## 场景描述

基于ModelArts Studio大模型即服务平台在模型广场预置的模型模板，用户可以使用推荐的模型权重文件或自定义的模型权重文件，创建一个自己的模型。

创建成功的模型可以在ModelArts Studio大模型即服务平台进行调优、压缩、推理等操作。

## 约束限制

- 用于生成专属模型的模型权重文件需要满足Hugging Face上的对应模型的文件格式要求。
  - 模型权重文件夹下包括权重类文件、词表类文件和配置类文件。
  - 可以使用transformers的from_pretrained方法对模型权重文件夹进行加载。

  具体请参见Hugging Face官方文档**Documentations**。
- 当选择ChatGLM3-6B、GLM-4-9B、Qwen-7B、Qwen-14B和Qwen-72B模型框架时（模型名字必须一致），需要修改权重配置才能正常运行模型，操作步骤请参见**修改权重配置**。

## 前提条件

已准备好用于生成专属模型的模型权重文件，并存放于OBS桶中，OBS桶必须和MaaS服务在同一个Region下。

## 修改权重配置

当选择ChatGLM3-6B、GLM-4-9B、Qwen-7B、Qwen-14B和Qwen-72B模型框架创建模型时（模型名字必须一致），如果使用自定义模型权重文件，则需要修改权重配置才能正常运行模型；如果使用推荐的模型权重文件，则不需要修改权重配置，可以跳过该步骤。修改后的权重文件要更新至OBS桶中。

- ChatGLM3-6B、GLM-4-9B

  修改文件“tokenization_chatglm.py”。

- 第一处

  原内容

  ```
  # Load from model defaults assert self.padding_side == "left"
  ```

  修改为

  ```
  # Load from model defaults # assert self.padding_side == "left"
  ```

- 第二处

  原内容

  ```
  if needs_to_be_padded:
  difference = max_length - len(required_input)
  if "attention_mask" in encoded_inputs:
  encoded_inputs["attention_mask"] = [0] * difference + encoded_inputs["attention_mask"]
  if "position_ids" in encoded_inputs:
  encoded_inputs["position_ids"] = [0] * difference + encoded_inputs["position_ids"]
  encoded_inputs[self.model_input_names[0]] = [self.pad_token_id] * difference + required_input
  ```

  修改为

  ```
  if needs_to_be_padded:
  difference = max_length - len(required_input)
  if "attention_mask" in encoded_inputs:
  encoded_inputs["attention_mask"] = encoded_inputs["attention_mask"] + [0] * difference
  if "position_ids" in encoded_inputs:
  encoded_inputs["position_ids"] = encoded_inputs["position_ids"] + [0] * difference
  encoded_inputs[self.model_input_names[0]] =  required_input + [self.pad_token_id] * difference
  ```

- Qwen-7B、Qwen-14B和Qwen-72B

  修改文件“modeling_qwen.py”。

  原内容

  ```
  SUPPORT_BF16 = SUPPORT_CUDA and torch.cuda.is_bf16_supported() SUPPORT_FP16 = SUPPORT_CUDA and torch.cuda.get_device_capability(0)[0] >= 7
  ```

  修改为

  ```
  SUPPORT_BF16 = SUPPORT_CUDA and True SUPPORT_FP16 = SUPPORT_CUDA and True
  ```

## 创建我的模型

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 进入创建模型页面。
   - 方式一：在ModelArts Studio左侧导航栏中，选择“我的模型”进入模型列表，单击“创建模型”弹出创建模型页面。
   - 方式二：在ModelArts Studio左侧导航栏中，选择“模型广场”，在模型广场选择模型并单击“立即使用”进入模型详情页，单击“创建模型”弹出创建模型页面。
4. 在创建模型页面，配置参数。

**表 5-1** 创建模型

| 参数 | 说明 |
| --- | --- |
| 来源模型 | • 当从“我的模型”进入创建模型页面时，单击选择基础模型完成模型选择。<br>• 当从“模型广场”进入创建模型页面时，此处默认呈现选择的模型。<br>当选择模型后，支持单击“重新选择”更改模型。 |
| 模型名称 | 自定义模型名称。<br>支持1~64位，以中文、大小写字母开头，只包含中文、大小写字母、数字、下划线（_）、中划线（-）和（.）。 |
| 描述 | 模型简介。支持100字符。 |
| 权重设置与词表 | 默认选择“使用推荐权重”，支持选择“自定义权重”。<br>• “使用推荐权重”：使用平台推荐的权重文件，可提高模型的训练、压缩、部署和调优等服务的使用效率。<br>• “自定义权重”：使用用户自定义的权重文件，需要先将权重文件上传至OBS桶中。且权重文件必须满足**约束限制**。<br>权重文件指的是模型的参数集合。<br>**说明**<br>百川系列模型只支持自定义权重。 |
| 选择自定义权重路径 | 当“权重设置与词表”选择“自定义权重”时，需要选择存放模型权重文件的OBS路径，必须选择到模型文件夹。 |
| 权重校验 | 当“权重设置与词表”选择“自定义权重”时，需要选择是否开启权重文件校验。默认是开启的。<br>• 当开启权重校验时，平台会对OBS中的权重文件进行校验，确认其是否满足规范。权限校验常见的失败情况及其处理建议请参见**权重校验**。<br>• 当关闭权重校验时，则不进行校验，创建模型可能会因为权重文件不合规而失败。 |

5. 参数配置完成后，单击“创建”，创建自定义模型。
   在模型列表，当模型“状态”变成“创建成功”时，表示模型创建完成。

## 查看我的模型详情

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“我的模型”进入模型列表。

4. 单击模型名称，进入模型详情页面，可以查看模型“基本信息”和“作业记录”。
   - 基本信息：可以查看模型名称、ID、来源模型等信息。
   - 作业记录：可以查看该模型被用于哪些作业类型，以及当前作业的状态等信息。

## 删除我的模型

**说明**

删除操作无法恢复，请谨慎操作。

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“我的模型”进入模型列表。
4. 在模型列表，单击模型名称，进入模型详情页面，查看模型的“作业记录”。
   - 如果作业记录为空，则直接执行下一步。
   - 如果作业记录存在作业，则先删除所有作业，再执行下一步。

   **说明**

   当模型存在作业记录会删除失败。
5. 在模型详情页，单击右上角的“删除”，在弹窗中输入“DELETE”，单击“确定”，删除模型。

## 权重校验

创建模型时，开启权重校验后，平台会自动创建一个权重校验的任务，在模型详情页的作业记录列表可以查看权重校验任务。

**图 5-1** 查看权重校验任务


当状态显示运行失败时，鼠标悬停在状态即可查看失败信息，根据失败信息处理问题。常见的权限校验失败信息及其处理建议请参见表5-2。

**表 5-2** 权重校验常见的失败信息

| 失败信息 | 信息解释 | 处理建议 |
| --- | --- | --- |
| Unknown error, please contact the operation and maintenance personnel or check the log to locate the specific problem. | 未知错误。 | 查看日志定位处理问题，或者联系技术支持。 |
| Backend model template selection error (metadata error). | 后台模型模板选择错误。 | 查看日志定位处理问题，或者联系技术支持。 |
| Failed to read standard config.json in the background. | 后台读取标准config.json失败。 | 查看日志定位处理问题，或者联系技术支持。 |
| Failed to read generation_config.json. | generation_config.json内容格式错误。 | 检查“generation_config.json”文件中的内容是否为json格式。 |
| The value of do_sample is not set to true in generation_config.json, which is inconsistent with the configured sampling parameters such as temperature, top_p, top_k etc. | 在generation_config.json中没有将do_sample的值设置为true，与配置的temperature、top_p、top_k等采样参数矛盾。 | 将“generation_config.json”文件中的“do_sample”的值设置为“true”。 |
| Failed to read user config.json. | config.json不存在或内容不符合json格式。 | 检查“config.json”文件是否存在，或者是内容是否为json格式。 |
| The quantization_config field is missing in config.json, please check whether it is awq quantization weight. | config.json中缺少quantization_config字段，请检查是否为awq量化权重。 | 检查权重和模型模板是否匹配。 |
| There is an extra quantization_config field in config.json. Please check whether it is a non-quantized weight. | config.json中多出quantization_config字段，请检查是否为非量化权重。 | 检查权重和模型模板是否匹配。 |
| Key fields describing the model structure are missing from config.json, or their values are inconsistent with standard open source. | config.json中缺少描述模型结构的关键字段，或其值与标准开源不一致。 | 检查“config.json”文件中的配置是否与模型官方一致。 |

| 失败信息 | 信息解释 | 处理建议 |
| --- | --- | --- |
| Error loading tokenizer in transformers. | tranformers加载tokenizer出错。 | 检查词表文件是否正确。 |
| Error loading weights in transformers. | tranformers加载权重出错。 | 检查权重文件是否正确。 |

# 6 使用 MaaS 调优模型

在ModelArts Studio大模型即服务平台完成模型创建后，可以对模型进行调优，获得更合适的模型。

## 场景描述

从“我的模型”中选择一个模型进行调优，当模型完成调优任务后会产生一个新的模型，呈现在“我的模型”列表中。

## 约束限制

表6-1列举了支持模型调优的模型，不在表格里的模型不支持使用MaaS调优模型。

**说明**

当选择ChatGLM3-6B、GLM-4-9B、Qwen-7B、Qwen-14B和Qwen-72B模型框架进行模型调优时，在创建模型时需要修改权重配置才能正常运行模型。详细配置请参见**修改权重配置**。

**表 6-1** 支持模型微调的模型

| 模型名称 | 全参微调 | lora微调 |
| --- | --- | --- |
| Baichuan2-13B | √ | √ |
| ChatGLM3-6B | √ | √ |
| Llama2-13B | √ | √ |
| Llama2-70B | √ | √ |
| Llama2-7B | √ | √ |
| Llama3-70B | √ | √ |
| Llama3-8B | √ | √ |
| Qwen1.5-14B | √ | √ |
| Qwen1.5-32B | √ | √ |
| Qwen1.5-72B | √ | √ |

| 模型名称 | 全参微调 | lora微调 |
| --- | --- | --- |
| Qwen1.5-7B | √ | √ |
| Qwen2-72B | √ | √ |
| Qwen2-72B-1K | √ | √ |
| Qwen2-7B | √ | √ |
| Qwen-72B | √ | √ |
| Qwen-14B | √ | √ |
| Qwen-7B | √ | √ |
| Qwen2-1.5B | √ | √ |
| Qwen2-0.5B | √ | √ |

## 支持的数据集格式

- **jsonl格式**

  一行数据就是数据集中的一条样本，建议总的数据样本不少于2000条，如下所示是一行数据集的示例，单轮对话也可以复用此格式。

  ```
  {"conversation_id": 1, "chat": {"turn_1": {"Human":"text","MOSS":"text"},"turn_2": {"Human":"text","MOSS":"text"}}}
  ```

  - “conversation_id”是样本编号。
  - “chat”后面是多轮对话的内容
  - “turn_n”表示是第n次对话，每次对话都有输入（对应Human角色）和输出（对应MOSS角色）。其中Human和MOSS仅用于角色区分，模型训练的内容只有text指代的文本。

  单击**下载**，获取示例数据集“simple_moss.jsonl”，该数据集可以用于文本生成类型的模型调优。

- **xlsx和csv格式**

  表格里的一行数据就是一条样本。表格中仅有3个字段：conversation_id、human和assistant。

  - conversation_id：对话ID，可以重复，但必须是正整数。若有多组Human-assiant对话使用同一个ID，则会按照文件中的顺序，将这几组对话编排成一个多轮对话。
  - human：对话输入，内容不能为空。
  - assistant：对话输出，内容不能为空。

  **说明**

  请按数据集格式要求准备数据，否则会导致调优任务失败。

## 前提条件

- 在“我的模型”页面存在已创建成功的模型。
- 已准备好训练数据集，并存放于OBS桶中，OBS桶必须和MaaS服务在同一个Region下。

- 当需要永久保存日志时，需要准备好存放日志的OBS路径，OBS桶必须和MaaS服务在同一个Region下。

## 创建调优任务

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型调优”进入任务列表。
4. 单击“创建调优任务”进入创建页面，完成创建配置。

**表 6-2** 创建调优任务

| 参数 | | 说明 |
| --- | --- | --- |
| 任务设置 | 任务名称 | 自定义调优任务名称。<br>支持1~64位，以中文、大小写字母开头，只包含中文、大小写字母、数字、中划线、下划线的名称。 |
| | 描述 | 调优任务简介。支持1000字符。 |
| 模型设置 | 模型来源 | 单击“选择模型”，从“我的模型”列表中选择需要调优的模型。 |
| | 选择调优类型 | • **全参微调：**直接在模型上训练，影响模型全量参数的微调训练，效果较好，收敛速度较慢，训练时间较长。<br>• **LoRA微调：**冻结原模型，通过往模型中加入额外的网络层，并只训练这些新增的网络层参数，效果接近或略差于全参训练，收敛速度快，训练时间短。 |
| | 调优后的模型名称 | 设置调优后产生的新模型的名称。<br>支持1~64位，以中文、大小写字母开头，只包含中文、大小写字母、数字、下划线（_）、中划线（-）和（.）。 |
| 数据设置 | 添加数据集 | 选择存放训练数据集的OBS路径，必须选择到文件。<br>**说明**<br>数据集必须满足要求（请参见**约束限制**），否则调优会失败。 |
| | 调优后模型权重保存路径 | 选择存放调优后的模型权重文件的OBS路径。<br>**说明**<br>权重文件要存放在空文件夹中，否则会覆盖原有文件。 |

| 参数 | | 说明 |
| --- | --- | --- |
| 超参设置 | 迭代步数/Iterations | 设置模型参数/权重更新的次数。在调优过程中，每一个Iterations会消耗32条训练数据。<br>当数据集是数百量级，则建议迭代4~8个epoch（epoch表示整个数据集被完整地用于一次训练的次数）；当数据集是数千量级，则建议迭代2~4个epoch；当数据集是更大数量，则建议迭代1~2个epoch。<br>总Iterations = 整个数据集完整训练需要的Iterations * epoch。例如，当一个数据集有3200条数据，完整训练一个数据集的Iterations为100，迭代2个epoch，总Iterations就是200。<br>取值范围：0~100000<br>默认值：1000 |
| | 学习率/learning_rate | 设置每个迭代步数（iteration）模型参数/权重更新的速率。学习率设置得过高会导致模型难以收敛，过低则会导致模型收敛速度过慢。<br>取值范围：0~0.1<br>默认值：0.00002<br>建议微调场景的学习率设置在$10^{-5}$这个量级。 |
| 资源设置 | 资源池类型 | 资源池分为公共资源池与专属资源池。<br>● 公共资源池供所有租户共享使用。<br>● 专属资源池需单独创建，不与其他租户共享。 |
| | 实例规格 | 选择实例规格，规格中描述了服务器类型、型号等信息，仅显示模型支持的资源 |
| | 实例数 | 设置实例数。 |
| 更多选项 | 永久保存日志 | 选择是否打开“永久保存日志”开关。<br>● 开关关闭（默认关闭）：表示不永久保存日志，则任务日志会在30天后会被清理。可以在任务详情页下载全部日志至本地。<br>● 开关打开：表示永久保存日志，此时必须配置“日志路径”，系统会将任务日志永久保存至指定的OBS路径。 |

| 参数 | | 说明 |
| --- | --- | --- |
| | 事件通知 | 选择是否打开“事件通知”开关。<br>● 开关关闭（默认关闭）：表示不启用消息通知服务。<br>● 开关打开：表示订阅消息通知服务，当任务发生特定事件（如任务状态变化或疑似卡死）时会发送通知。此时必须配置“主题名”和“事件”。<br>– “主题名”：事件通知的主题名称。单击“创建主题”，前往消息通知服务中创建主题。<br>– “事件”：选择要订阅的事件类型。例如“创建中”、“已完成”、“运行失败”等。<br>**说明**<br>● 需要为消息通知服务中创建的主题添加订阅，当订阅状态为“已确认”后，方可收到事件通知。订阅主题的详细操作请参见**添加订阅**。<br>● 使用消息通知服务会产生相关服务费用，详细信息请参见**计费说明**。 |
| | 自动停止 | 当使用付费资源时，可以选择是否打开“自动停止”开关。<br>● 开关关闭（默认关闭）：表示任务将一直运行直至完成。<br>● 开关打开：表示启用自动停止功能，此时必须配置自动停止时间，支持设置为“1小时”、“2小时”、“4小时”、6小时或“自定义”。启用该参数并设置时间后，运行时长到期后将会自动终止任务，准备排队等状态不扣除运行时长。 |

5. 参数配置完成后，单击“提交”，创建调优任务。
   在任务列表，当模型“状态”变成“已完成”时，表示模型调优完成。

## 模型调优时长估算

调优时长表示调优任务的“状态”处于“运行中”的耗时。由于训练吞吐有上下限，因此计算出的调优时长是个区间。

- 计算公式：调优时长 = 经验系数 x Iterations ÷（卡数 x 实例数 x 吞吐）
- 单位：小时

**表 6-3** 参数说明

| 参数 | 说明 |
| --- | --- |
| 经验系数 | 经验系数与模型训练迭代过程中处理的序列长度和批次大小有关，当前默认为36。 |

| 参数 | 说明 |
| --- | --- |
| Iterations | 创建调优任务时设置的“迭代步数/Iterations”超参值。 |
| 卡数 | 和创建调优任务时选择的“实例规格”相关，例如，“实例规格”选择的是“Ascend: 2*ascend-snt9b2(64GB)”，*号前面的数字是2，则卡数就是2。 |
| 实例数 | 创建调优任务时设置的“实例数”。 |
| 吞吐 | 吞吐表示每秒每卡处理的Tokens数量，吞吐值的上下限可以参考表6-4获取。<br>单位：tokens/s/p |

**表 6-4** 各模型的吞吐数据参考

| 模型名称 | 训练类型 | 吞吐下限取整 | 吞吐上限取整 |
| --- | --- | --- | --- |
| Baichuan2-13B | sft | 1200 | 1600 |
| | lora | 1300 | 1800 |
| ChatGLM3-6B | sft | 2000 | 2700 |
| | lora | 2300 | 3100 |
| Llama2-13B | sft | 1300 | 1800 |
| | lora | 1400 | 1900 |
| Llama2-70B | sft | 300 | 400 |
| | lora | 400 | 500 |
| Llama2-7B | sft | 3100 | 4200 |
| | lora | 3500 | 4700 |
| Llama3-70B | sft | 300 | 400 |
| | lora | 300 | 500 |
| Llama3-8B | sft | 2100 | 2800 |
| | lora | 2300 | 3100 |
| Qwen-14B | sft | 1200 | 1600 |
| | lora | 1400 | 1900 |
| Qwen-72B | sft | 300 | 400 |
| | lora | 300 | 500 |
| Qwen-7B | sft | 2100 | 2900 |
| | lora | 2200 | 3000 |

| 模型名称 | 训练类型 | 吞吐下限取整 | 吞吐上限取整 |
| --- | --- | --- | --- |
| Qwen1.5-14B | sft | 1300 | 1700 |
| | lora | 1400 | 1800 |
| Qwen1.5-32B | sft | 600 | 800 |
| | lora | 700 | 900 |
| Qwen1.5-72B | sft | 300 | 400 |
| | lora | 300 | 500 |
| Qwen1.5-7B | sft | 2200 | 3000 |
| | lora | 2600 | 3600 |
| Qwen2-0.5B | sft | 12800 | 17300 |
| | lora | 12800 | 17300 |
| Qwen2-1.5B | sft | 7300 | 9800 |
| | lora | 7300 | 9900 |
| Qwen2-72B | sft | 300 | 300 |
| | lora | 300 | 400 |
| Qwen2-72B-1K | sft | 300 | 300 |
| | lora | 300 | 400 |
| Qwen2-7B | sft | 2300 | 3200 |
| | lora | 2600 | 3500 |

## 查看调优任务详情

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型调优”进入任务列表。
4. 单击任务名称，进入调优任务详情页面，可以查看任务详情和日志。
   - “详情”：可以查看任务的基本信息，包括任务、模型、资源等设置信息。
   - “日志”：可以搜索、查看和下载任务日志。
     - 查看loss：当作业进入训练流程之后，会按照Step进行loss打印，因此在日志中搜索关键字段“lm loss”即可查看loss。
     - 获取训练吞吐数据：在打印的loss日志中搜索关键字段“elapsed time per iteration”获取每步迭代耗时，总的Token数可以用日志中的“global batch size”和“SEQ_LEN”相乘获得，训练的每卡每秒的吞吐=总Token数÷每步迭代耗时÷总卡数。

## 删除调优任务

> **说明**
>
> 删除操作无法恢复，请谨慎操作。

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型调优”进入任务列表。
4. 选择调优任务，单击操作列的“删除”，在弹窗中输入“DELETE”，单击“确定”，删除任务。

# 7 使用 MaaS 压缩模型

在ModelArts Studio大模型即服务平台完成模型创建后，可以对模型进行压缩，获得更合适的模型。

## 场景描述

模型压缩是指将高比特浮点数映射到低比特量化空间，从而减少显存占用的资源，降低推理服务时延，提高推理服务吞吐量，并同时减少模型的精度损失。模型压缩适用于追求更高的推理服务性能、低成本部署以及可接受一定精度损失的场景。

ModelArts Studio大模型即服务平台当前支持SmoothQuant-W8A8和AWQ-W4A16两种压缩策略。

**表 7-1** 压缩策略的适用场景

| 压缩策略 | 场景 |
| --- | --- |
| SmoothQuant-W8A8 | • 长序列的场景<br>• 大并发量的场景 |
| AWQ-W4A16 | • 小并发量的低时延场景<br>• 更少推理卡数部署的场景 |

## 约束限制

表7-2列举了支持模型压缩的模型，不在表格里的模型不支持使用MaaS压缩模型。

**表 7-2** 支持模型压缩的模型

| 模型名称 | SmoothQuant-W8A8 | AWQ-W4A16 |
| --- | --- | --- |
| Llama2-13B | √ | √ |
| Llama2-70B | √ | √ |
| Llama2-7B | √ | √ |

| 模型名称 | SmoothQuant-W8A8 | AWQ-W4A16 |
| --- | --- | --- |
| Llama3-70B | √ | √ |
| Llama3-8B | √ | √ |
| Qwen1.5-14B | √ | √ |
| Qwen1.5-72B | √ | √ |
| Qwen1.5-7B | √ | √ |
| Qwen2-72B | √ | x |
| Qwen2-72B-1K | √ | x |

## 前提条件

- 在“我的模型”页面存在已创建成功的模型。
- 已准备好用于存放压缩后模型权重文件的OBS桶，OBS桶必须和MaaS服务在同一个Region下。

## 创建压缩任务

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型压缩”进入任务列表。
4. 单击“创建压缩任务”进入创建页面，完成创建配置。

**表 7-3** 创建压缩任务

| 参数 | | 说明 |
| --- | --- | --- |
| 任务设置 | 任务名称 | 自定义压缩任务名称。<br>支持1~64位，以中文、大小写字母开头，只包含中文、大小写字母、数字、中划线、下划线的名称。 |
| | 描述 | 压缩任务简介。支持1000字符。 |
| 模型设置 | 模型来源 | 单击“选择模型”，从“我的模型”列表中选择需要压缩的模型。 |

| 参数 | | 说明 |
| --- | --- | --- |
| | 压缩策略 | • **SmoothQuant-W8A8：**SmoothQuant是一种同时确保准确率与推理高效的训练后量化（PTQ）方法，W8A8可实现8-bit权重、8-bit激活（W8A8）量化，引入平滑因子来平滑激活异常值，将量化难度从较难量化的激活转移到容易量化的权重上。<br>• **AWQ-W4A16：**AWQ是一种大模型低比特权重的训练后量化（PTQ）方法，W4A16可实现4-bit权重、16-bit激活（W4A16）量化，通过激活值来选择并放大显著权重，以提高推理效率。 |
| | 压缩后模型名称 | 设置压缩后产生的新模型的名称。<br>支持1~64位，以中文、大小写字母开头，只包含中文、大小写字母、数字、下划线（_）、中划线（-）和（.）。 |
| 参数设置 | 平滑系数/Migration Strength | 设置SmoothQuant量化的迁移系数，仅“压缩策略”选择“SmoothQuant-W8A8”时才需要配置。建议使用默认值。<br>取值范围：0~1<br>默认值：0.5 |
| | 压缩后模型权重保存路径 | 选择压缩后模型权重文件存放的OBS路径。 |
| 资源设置 | 资源池类型 | 资源池分为公共资源池与专属资源池。<br>• 公共资源池供所有租户共享使用。<br>• 专属资源池需单独创建，不与其他租户共享。 |
| | 实例规格 | 选择实例规格，规格中描述了服务器类型、型号等信息。 |
| 更多选项 | 永久保存日志 | 选择是否打开“永久保存日志”开关。<br>• 开关关闭（默认关闭）：表示不永久保存日志，则任务日志会在30天后会被清理。可以在任务详情页下载全部日志至本地。<br>• 开关打开：表示永久保存日志，此时必须配置“日志路径”，系统会将任务日志永久保存至指定的OBS路径。 |

| 参数 | | 说明 |
| --- | --- | --- |
| | 事件通知 | 选择是否打开“事件通知”开关。<br>• 开关关闭（默认关闭）：表示不启用消息通知服务。<br>• 开关打开：表示订阅消息通知服务，当任务发生特定事件（如任务状态变化或疑似卡死）时会发送通知。此时必须配置“主题名”和“事件”。<br>– “主题名”：事件通知的主题名称。单击“创建主题”，前往消息通知服务中创建主题。<br>– “事件”：选择要订阅的事件类型。例如“创建中”、“已完成”、“运行失败”等。<br>**说明**<br>• 需要为消息通知服务中创建的主题添加订阅，当订阅状态为“已确认”后，方可收到事件通知。订阅主题的详细操作请参见**添加订阅**。<br>• 使用消息通知服务会产生相关服务费用，详细信息请参见**计费说明**。 |
| | 自动停止 | 当使用付费资源时，可以选择是否打开“自动停止”开关。<br>• 开关关闭（默认关闭）：表示任务将一直运行直至完成。<br>• 开关打开：表示启用自动停止功能，此时必须配置自动停止时间，支持设置为“1小时”、“2小时”、“4小时”、6小时或“自定义”。启用该参数并设置时间后，运行时长到期后将会自动终止任务，准备排队等状态不扣除运行时长。 |

5. 参数配置完成后，单击“提交”，创建压缩任务。
   在任务列表，当模型“状态”变成“已完成”时，表示模型压缩完成。

## 模型压缩时长估算

**表 7-4** 模型压缩时长估算

| 模型名称 | SmoothQuant-W8A8 | AWQ-W4A16 |
| --- | --- | --- |
| Llama2-13B | 10~20分钟 | 60分钟 |
| Llama2-70B | 40分钟 | 3小时 |
| Llama2-7B | 10~20分钟 | 40分钟 |
| Llama3-70B | 40分钟 | 3小时 |

| 模型名称 | SmoothQuant-W8A8 | AWQ-W4A16 |
| --- | --- | --- |
| Llama3-8B | 10~20分钟 | 40分钟 |
| Qwen1.5-14B | 10~20分钟 | 60分钟 |
| Qwen1.5-72B | 40分钟 | 3小时 |
| Qwen1.5-7B | 10~20分钟 | 40分钟 |
| Qwen2-72B | 40分钟 | - |
| Qwen2-72B-1K | 40分钟 | - |

## 查看压缩任务信息

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型压缩”进入任务列表。
4. 单击任务名称，进入压缩任务详情页面，可以查看任务详情和日志。
   - “详情”：可以查看任务的基本信息，包括任务、模型、资源等设置信息。
   - “日志”：可以搜索、查看和下载任务日志。

## 删除压缩任务

**说明**

删除操作无法恢复，请谨慎操作。

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型压缩”进入任务列表。
4. 选择压缩任务，单击操作列的“删除”，在弹窗中输入“DELETE”，单击“确定”，删除任务。

# 8 使用 MaaS 部署模型服务

在ModelArts Studio大模型即服务平台可以将模型部署为服务，便于在“模型体验”或其他业务环境中可以调用。

## 约束限制

部署模型服务时，ModelArts Studio大模型即服务平台预置了推理的最大输入输出长度。模型Qwen-14B默认是2048，模型Qwen2-72B-32K默认是32768，其他模型默认都是4096。

## 前提条件

在“我的模型”页面存在已创建成功的模型。

## 部署模型服务

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型部署”进入服务列表。
4. 单击“部署模型服务”进入部署页面，完成创建配置。

**表 8-1** 部署模型服务

| 参数 | | 说明 |
| --- | --- | --- |
| 服务设置 | 服务名称 | 自定义部署模型服务的名称。<br>支持1~64位，以中文、大小写字母开头，只包含中文、大小写字母、数字、中划线、下划线的名称。 |
| | 描述 | 部署模型服务的简介。支持256字符。 |
| 模型设置 | 部署模型 | 单击“选择模型”，从“我的模型”列表中选择需要部署的模型。 |

| 参数 | | 说明 |
| --- | --- | --- |
| 资源设置 | 资源池类型 | 资源池分为公共资源池与专属资源池。<br>● 公共资源池供所有租户共享使用。<br>● 专属资源池需单独创建，不与其他租户共享。 |
| | 实例规格 | 选择实例规格，规格中描述了服务器类型、型号等信息。 |
| | 单实例流量限制（QPS） | 设置单实例的QPS，可以参考**QPS的推荐值说明**设置待部署模型的QPS值。<br>单位：次/秒<br>**说明**<br>在部署过程中出现错误码“ModelArts.4206”时，表示QPS请求数量达到限制，建议等待限流结束后再重启服务。 |
| | 实例数 | 设置服务器个数。设置多个实例可提高总QPS，“总QPS=单实例QPS x 实例数”。 |
| 更多选项 | 事件通知 | 选择是否打开“事件通知”开关。<br>● 开关关闭（默认关闭）：表示不启用消息通知服务。<br>● 开关打开：表示订阅消息通知服务，当任务发生特定事件（如任务状态变化或疑似卡死）时会发送通知。此时必须配置“主题名”和“事件”。<br>– “主题名”：事件通知的主题名称。单击“创建主题”，前往消息通知服务中创建主题。<br>– “事件”：选择要订阅的事件类型。例如“创建中”、“已完成”、“运行失败”等。<br>**说明**<br>● 需要为消息通知服务中创建的主题添加订阅，当订阅状态为“已确认”后，方可收到事件通知。订阅主题的详细操作请参见**添加订阅**。<br>● 使用消息通知服务会产生相关服务费用，详细信息请参见**计费说明**。 |
| | 自动停止 | 当使用付费资源时，可以选择是否打开“自动停止”开关。<br>● 开关关闭（默认关闭）：表示任务将一直运行。<br>● 开关打开：表示启用自动停止功能，此时必须配置自动停止时间，支持设置为“1小时”、“2小时”、“4小时”、6小时或“自定义”。启用该参数并设置时间后，运行时长到期后将会自动终止任务，准备排队等状态不扣除运行时长。 |

5. 参数配置完成后，单击“提交”，创建部署任务。
   在任务列表，当模型“状态”变成“运行中”时，表示模型部署完成。

## 查看部署任务信息

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型部署”进入服务列表。
4. 单击服务名称，进入部署模型服务详情页面，可以查看服务信息。
   - “详情”：可以查看服务的基本信息，包括服务、模型、资源等设置信息。
   - “监控”：可以查看服务监控和资源监控信息。

     **说明**

     “算力利用率”表示每分钟NPU的平均使用率，当请求率较低时，使用率会显示为0。
   - “事件”：可以查看服务的事件信息。事件保存周期为1个月，1个月后自动清理数据。
   - “日志”：可以搜索和查看服务日志。

## 删除部署任务

**说明**

删除操作无法恢复，请谨慎操作。

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型部署”进入服务列表。
4. 选择部署模型服务，单击操作列的“更多 > 删除”，在弹窗中输入“DELETE”，单击“确定”，删除服务。

## QPS 的推荐值说明

单实例流量限制QPS和请求的输入输出有关，**表8-2**中的QPS推荐值是在多轮对话、摘要生产和信息检索场景下预估出的数据，仅供参考，如果要了解其余典型场景的QPS推荐值请联系技术支持。

单位：次/秒

**表 8-2** 各模型的 QPS 推荐值

| 模型名称 | QPS推荐值 |
| --- | --- |
| Baichuan2-13B | 1 |
| ChatGLM3-6B | 3 |
| Llama2-13B | 1 |

| 模型名称 | QPS推荐值 |
| --- | --- |
| Llama2-13B-AWQ | 1 |
| Llama2-13B-SQ | 1 |
| Llama2-70B | 1 |
| Llama2-70B-AWQ | 1 |
| Llama2-70B-SQ | 1 |
| Llama2-7B | 3 |
| Llama2-7B-AWQ | 3 |
| Llama2-7B-SQ | 3 |
| Llama3-70B | 1 |
| Llama3-70B-AWQ | 1 |
| Llama3-70B-SQ | 1 |
| Llama3-8B | 3 |
| Llama3-8B-AWQ | 3 |
| Llama3-8B-SQ | 6 |
| Llama3.1-70B | 1 |
| Llama3.1-8B | 3 |
| Qwen1.5-14B | 1 |
| Qwen1.5-14B-AWQ | 1 |
| Qwen1.5-14B-SQ | 1 |
| Qwen1.5-32B | 1 |
| Qwen1.5-72B | 1 |
| Qwen1.5-72B-AWQ | 1 |
| Qwen1.5-72B-SQ | 1 |
| Qwen1.5-7B | 3 |
| Qwen1.5-7B-AWQ | 3 |
| Qwen1.5-7B-SQ | 3 |
| Qwen-14B | 1 |
| Qwen2-72B | 1 |
| Qwen2-72B-AWQ | 1 |
| Qwen2-72B-SQ | 1 |

| 模型名称 | QPS推荐值 |
| --- | --- |
| Qwen2-72B-1K | 1 |
| Qwen2-72B-32K | 1 |
| Qwen2-7B | 3 |
| Qwen2-7B-AWQ | 3 |
| Qwen-72B | 1 |
| Qwen-7B | 3 |
| Qwen2-1.5B | 6 |
| Qwen2-0.5B | 9 |
| Glm-4-9B | 3 |
| Yi-34B | 1 |
| Yi-6B | 3 |

# 9 调用 MaaS 部署的模型服务

在ModelArts Studio大模型即服务平台部署成功的模型服务支持在其他业务环境中调用。

## 约束限制

只有“状态”是“运行中”的模型服务才支持被调用。

## 步骤 1：获取 API Key

在调用MaaS部署的模型服务时，需要填写API Key用于接口的鉴权认证。

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“鉴权管理”。
4. 在“鉴权管理”页面，单击“创建API Key”，填写描述信息后，单击“确认”会返回“您的密钥”，请复制保存密钥，单击“关闭”后将无法再次查看密钥。

   **说明**

   - 最多支持创建5个密钥，密钥只会在新建后显示一次，请妥善保存。
   - 当密钥丢失将无法找回，请新建API Key获取新的访问密钥。

## 步骤 2：调用 MaaS 模型服务进行预测

1. 在ModelArts Studio左侧导航栏中，选择“模型部署”进入服务列表。
2. 选择要调用的服务，单击操作列的“更多 > 调用”，复制Python脚本用于业务环境调用。

   示例代码如下所示：

```
# coding=utf-8

import requests
import json

if __name__ == '__main__':
    url = "xxxxxxxxxx/v1/chat/completions"

    # Send request.
    headers = {
```

```
        'Content-Type': 'application/json',
        'Authorization': 'Bearer yourApiKey'  # 把yourApiKey替换成已获取的API Key。例如，获取的API Key是“1234abcd...”时，此处填写“Bearer 1234abcd...”。
    }
    data = {
        "model": "Qwen2-7B",  # 调用时的模型名称
        "max_tokens": 20,
        "messages": [
            {"role": "system", "content": "You are a helpful assistant."},
            {"role": "user", "content": "hello"}
        ]
    }
    resp = requests.post(url, headers=headers, data=json.dumps(data), verify=False)

    # Print result.
    print(resp.status_code)
    print(resp.text)
```

模型服务的API与vLLM相同，表9-1仅介绍关键参数，详细参数解释请参见vLLM官网Sampling Parameters。

**表 9-1** 请求参数说明

| 参数 | 是否必选 | 默认值 | 参数类型 | 描述 |
| --- | --- | --- | --- | --- |
| model | 是 | 无 | Str | 调用时的模型名称。<br>在ModelArts Studio大模型即服务平台的“模型部署”页面，选择调用的模型服务，单击操作列的“更多 > 调用”，在调用页面可以获取“模型名称”。 |
| messages | 是 | - | Array | 请求输入的问题。 |
| max_tokens | 否 | 16 | Int | 每个输出序列要生成的最大Tokens数量。 |
| top_k | 否 | -1 | Int | 控制要考虑的前几个Tokens的数量的整数。设置为“-1”表示考虑所有Tokens。<br>适当降低该值可以减少采样时间。 |
| top_p | 否 | 1.0 | Float | 控制要考虑的前几个Tokens的累积概率的浮点数。<br>取值范围：0~1<br>设置为“1”表示考虑所有Tokens。 |
| temperature | 否 | 1.0 | Float | 控制采样的随机性的浮点数。较低的值使模型更加确定性，较高的值使模型更加随机。“0”表示贪婪采样。 |
| stop | 否 | None | None/Str/List | 用于停止生成的字符串列表。返回的输出将不包含停止字符串。<br>例如，设置为["你"，"好"]时，在生成文本过程中，遇到“你”或者“好”将停止文本生成。 |

| 参数 | 是否必选 | 默认值 | 参数类型 | 描述 |
| --- | --- | --- | --- | --- |
| stream | 否 | False | Bool | 是否开启流式推理。默认为“False”，表示不开启流式推理。 |
| n | 否 | 1 | Int | 返回多条正常结果。<br>● 不使用beam_search场景下，n取值建议为1≤n≤10。如果n>1时，必须确保不使用greedy_sample采样，也就是top_k > 1，temperature > 0。<br>● 使用beam_search场景下，n取值建议为1<n≤10。如果n=1，会导致推理请求失败。<br>**说明**<br>n建议取值不超过10，n值过大会导致性能劣化，显存不足时，推理请求会失败。 |
| use_beam_search | 否 | False | Bool | 是否使用beam_search替换采样。<br>使用该参数时，如下参数必须按要求设置。<br>● n：大于1<br>● top_p：1.0<br>● top_k：-1<br>● temperature：0.0 |
| presence_penalty | 否 | 0.0 | Float | presence_penalty表示会根据当前生成的文本中新出现的词语进行奖惩。取值范围[-2.0,2.0]。 |
| frequency_penalty | 否 | 0.0 | Float | frequency_penalty会根据当前生成的文本中各个词语的出现频率进行奖惩。取值范围[-2.0,2.0]。 |
| length_penalty | 否 | 1.0 | Float | length_penalty表示在beam search过程中，对于较长的序列，模型会给予较大的惩罚。<br>使用该参数时，必须添加如下三个参数，且必须按要求设置。<br>● top_k：-1<br>● use_beam_search：true<br>● best_of：大于1 |
| ignore_eos | 否 | False | Bool | ignore_eos表示是否忽略EOS并且继续生成Token。 |

返回示例如下所示。

```
{
    "id": "cmpl-29f7a172056541449eb1f9d31cfac162",
```

```
  "object": "chat.completion",
  "created": 1723190150,
  "model": "Qwen2-7B",
  "choices": [
     {
        "index": 0,
        "message": {
           "role": "assistant",
           "content": "你好！很高兴能为你提供帮助。有什么问题我可以回答或帮你解决吗？"
        },
        "logprobs": null,
        "finish_reason": "stop",
        "stop_reason": null
     }
  ],
  "usage": {
     "prompt_tokens": 20,
     "total_tokens": 38,
     "completion_tokens": 18
  }
}
```


**表 9-2** 返回参数说明

| 参数 | 参数类型 | 描述 |
| --- | --- | --- |
| id | Str | 请求ID。 |
| object | Str | 请求任务。 |
| created | Int | 请求生成的时间戳。 |
| model | Str | 调用的模型名。 |
| choices | Array | 模型生成内容。 |
| usage | Object | 请求输入长度、输出长度和总长度。 |

当调用失败时，可以根据错误码调整脚本或运行环境。

**表 9-3** 常见错误码

| 错误码 | 错误内容 | 说明 |
| --- | --- | --- |
| 400 | Bad Request | 请求包含语法错误。 |
| 403 | Forbidden | 服务器拒绝执行。 |
| 404 | Not Found | 服务器找不到请求的网页。 |
| 500 | Internal Server Error | 服务内部错误。 |

# 10 更新 MaaS 模型服务的模型权重

## 场景描述

用户在运用大型模型进行推理任务时，需定期对模型进行迭代和优化。为适应模型权重的更新和迭代，必须对已部署的服务执行相应的升级操作，以确保服务使用的是最新模型。

ModelArts Studio大模型即服务平台支持滚动升级模型权重，允许模型服务在运行时进行权重的迭代升级，该操作不会影响部署服务的正常运行。滚动升级模型权重的功能避免了重新部署整个模型服务的必要性，从而确保了服务的连续性，无需执行任何业务迁移操作。

## 约束限制

- 模型权重更新后，后续对部署模型进行操作，即从“我的模型”中对该部署模型发起的操作时，都将基于新权重进行。
- 仅当模型服务处于这几个状态下才能更新权重：运行中、异常、告警、停止。

## 步骤 1：验证模型权重文件

在进行模型服务升级之前，必须先确认模型权重文件能够成功完成推理任务。只有当验证成功，确保了模型权重的功能性和准确性后，才可以进行模型权重的滚动升级。

1. 获取待更新的模型权重文件，并上传到OBS桶中。
2. 参考**创建我的模型**，用待更新的模型权重文件新建一个我的模型。关键参数请参见**表10-1**。

**表 10-1** 创建模型的关键参数说明

| 参数 | 说明 |
| --- | --- |
| 来源模型 | 选择和待升级的模型服务的“部署模型”同一个模型框架。 |
| 权重设置与词表 | 选择“自定义权重”。 |
| 选择自定义权重路径 | 选择存放待更新的模型权重文件的OBS路径，必须选择到模型文件夹。 |

| 参数 | 说明 |
| --- | --- |
| 权重校验 | 开启权重文件校验。 |

3. 参考**部署模型服务**，用新建的模型部署模型服务。
   - “模型设置”选择上一步新建的模型。
   - “资源设置”和待升级的模型服务保持一致。
   - 其他参数自定义。
4. 参考**调用MaaS部署的模型服务**，用上一步部署的模型服务验证推理效果。
   - 如果推理结果正确，则使用该模型权重完成执行**步骤2：滚动升级模型权重**。
   - 如果推理结果不正确，请先排查原因，待能正常完成推理任务后再用该模型权重文件升级。

## 步骤 2：滚动升级模型权重

当模型权重文件验证成功后，可以开始模型权重的滚动升级。

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“模型部署”进入服务列表。
4. 在服务列表，选择要升级的服务，单击操作列的“更多 > 更新权重”。
5. 在更新权重弹窗中，设置“自定义权重上传路径”，选择验证成功的模型权重文件存放的OBS路径，必须选择到模型文件夹。
6. 设置完成后，单击“确定”，在“权重变更确认”弹窗中单击“确定”，开始更新权重。服务状态变成“升级中”。

# 11 在 MaaS 应用实践中心查看应用解决方案

ModelArts Studio大模型即服务平台提供了MaaS应用实践中心，为具体的应用场景提供一整套解决方案。

## 应用中心介绍

“MaaS应用实践中心”提供基于行业客户应用场景的AI解决方案。MaaS提供的模型服务和华为云各AI应用层构建工具之间相互连通，通过灵活的组合方案，来帮助客户快速解决模型落地应用时所面临的业务及技术挑战。

MaaS应用实践中心结合KooSearch企业搜索服务、盘古数字人大脑和Dify，为具体的客户应用场景提供一整套解决方案。

- KooSearch企业搜索服务：基于在MaaS开源大模型部署的模型API，搭建企业专属方案、LLM驱动的语义搜索、多模态搜索增强。
- 盘古数字人大脑：基于在MaaS开源大模型部署的模型API，升级智能对话解决方案，含智能客服、数字人。
- Dify：支持自部署的应用构建开源解决方案，用于Agent编排、自定义工作流。

## 操作步骤

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio左侧导航栏中，选择“应用实践中心”跳转到“MaaS应用实践中心”页面。
4. 在“行业解决方案”选择应用，跳转到应用详情页，了解应用实现方案。