ModelArts 2.0.9

产品介绍

文档版本	01
发布日期	2020-11-17

华为技术有限公司

商标声明


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

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

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。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

目 录

1 什么是 ModelArts..... 1
2 功能介绍..... 4
3 基础知识..... 5
3.1 AI 开发基本流程介绍..... 5
3.2 AI 开发基本概念..... 6
3.3 ModelArts 中常用概念..... 8
3.4 数据管理..... 9
3.5 开发环境..... 10
3.6 模型训练..... 11
3.7 模型部署..... 12
3.8 自动学习..... 13
4 ModelArts 支持的 AI 框架..... 15
5 与其他服务的关系..... 19
6 如何访问 ModelArts..... 21
7 计费说明..... 22
8 权限管理..... 25
9 配额说明..... 27

1 什么是 ModelArts

ModelArts是面向AI开发者的一站式开发平台，提供海量数据预处理及半自动化标注、大规模分布式训练、自动化模型生成及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期AI工作流。

“一站式”是指AI开发的各个环节，包括数据处理、算法开发、模型训练、模型部署都可以在ModelArts上完成。从技术上看，ModelArts底层支持各种异构计算资源，开发者可以根据需要灵活选择使用，而不需要关心底层的技术。同时，ModelArts支持Tensorflow、MXNet等主流开源的AI开发框架，也支持开发者使用自研的算法框架，匹配您的使用习惯。

ModelArts的理念就是让AI开发变得更简单、更方便。

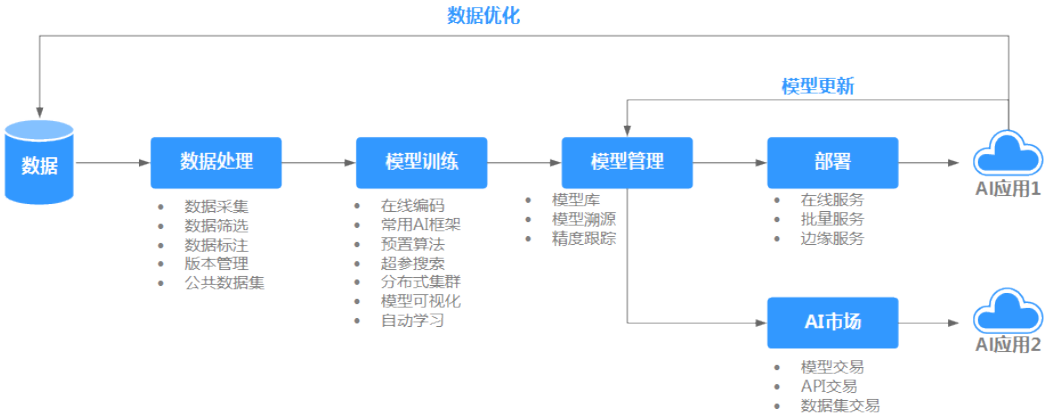
面向不同经验的AI开发者，提供便捷易用的使用流程。例如，面向业务开发者，不需关注模型或编码，可使用自动学习流程快速构建AI应用；面向AI初学者，不需关注模型开发，使用预置算法构建AI应用；面向AI工程师，提供多种开发环境，多种操作流程和模式，方便开发者编码扩展，快速构建模型及应用。

产品架构

ModelArts是一个一站式的开发平台，能够支撑开发者从数据到AI应用的全流程开发过程。包含数据处理、模型训练、模型管理、模型部署等操作，并且提供AI市场功能，能够在市场内与其他开发者分享模型。

ModelArts支持应用到图像分类、物体检测、视频分析、语音识别、产品推荐、异常检测等多种AI应用场景。

图 1-1 ModelArts 架构



产品优势

- **一站式**

  开“箱”即用，涵盖AI开发全流程，包含数据处理、模型开发、训练、管理、部署功能，可灵活使用其中一个或多个功能。

- **易上手**
  - 提供多种预置模型，开源模型想用就用。
  - 模型超参自动优化，简单快速。
  - 零代码开发，简单操作训练出自己的模型。
  - 支持模型一键部署到云、边、端。
- **高性能**
  - 自研MoXing深度学习框架，提升算法开发效率和训练速度。
  - 优化深度模型推理中GPU的利用率，加速云端在线推理。
  - 可生成在Ascend芯片上运行的模型，实现高效端边推理。
- **灵活**
  - 支持多种主流开源框架(TensorFlow、Spark_MLlib、MXNet、Caffe、PyTorch、XGBoost-Sklearn、MindSpore)。
  - 支持主流GPU和自研Ascend芯片。
  - 支持专属资源独享使用。
  - 支持自定义镜像满足自定义框架及算子需求。

首次使用 ModelArts

如果您是首次使用ModelArts的用户，建议您学习并了解如下信息：

- **基础知识了解**

  通过**基础知识**章节的内容，了解ModelArts相关的基础知识，包含AI开发的基础流程、AI开发的基础概念，以及ModelArts服务的特有概念和功能的详细介绍。

- **入门使用**

  针对不同角色的用户，您可以参考《**快速入门**》学习并上手使用ModelArts。《**快速入门**》提供了样例的详细操作指导，您可以基于此操作指导，在ModelArts服务中，构建一个模型或服务。

- **获取并尝试更多样例**

  ModelArts支持多种开源引擎，基于各类引擎和功能，提供了丰富的样例指导，您可以参考《**最佳实践**》的样例指导，完成相关的模型构建和部署。

- **使用更多的功能，并查看其相关操作指导**
  - 如果您是一个业务开发者，可以使用自动学习功能（无需编码，无需专业的AI基础能力），快速构建模型。详细操作指导可参考《**自动学习用户指南**》。
  - 如果您是一个AI初学者，可以使用一些常见的AI算法快速构建模型，无需编码开发模型。ModelArts基于常用的AI引擎内置了算法，您可以使用此预置算法快速构建模型。详细操作指导可参考《**AI工程师用户指南**》。
  - 如果您是一个AI工程师，可以使用AI全流程开发，包含数据管理、模型开发、训练、管理和部署等功能，您使用一个或多个功能应用到您的AI开发中。详细操作指导可参考《**AI工程师用户指南**》。
  - 如果您是一个开发者，想要直接调用ModelArts的API或SDK完成AI开发，您可以参考《**API参考**》或《**SDK参考**》获取详情。

2 功能介绍

繁多的AI工具安装配置、数据准备、模型训练慢等是困扰AI工程师的诸多难题。为解决这个难题，将一站式的AI开发平台（ModelArts）提供给开发者，从数据准备到算法开发、模型训练，最后把模型部署起来，集成到生产环境。一站式完成所有任务。

图 2-1 功能总览

ModelArts特色功能如下所示：

- **数据治理**

  支持数据筛选、标注等数据处理，提供数据集版本管理，特别是深度学习的大数据集，让训练结果可重现。

- **极“快”致“简”模型训练**

  自研的MoXing深度学习框架，更高效更易用，大大提升训练速度。

- **云边端多场景部署**

  支持模型部署到多种生产环境，可部署为云端在线推理和批量推理，也可以直接部署到端和边。

- **自动学习**

  支持多种自动学习能力，通过“自动学习”训练模型，用户不需编写代码即可完成自动建模、一键部署。

- **AI市场**

  预置常用算法和常用数据集，支持模型在企业内部共享或者公开共享。

3 基础知识

3.1 AI 开发基本流程介绍

什么是 AI

AI（人工智能）是通过机器来模拟人类认识能力的一种科技能力。AI最核心的能力就是根据给定的输入做出判断或预测。

AI 开发的目的是什么

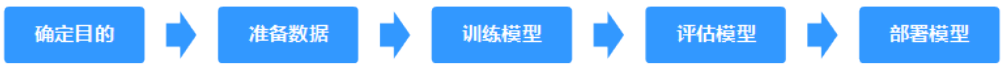
AI开发的目的是将隐藏在一大批数据背后的信息集中处理并进行提炼，从而总结得到研究对象的内在规律。

对数据进行分析，一般通过使用适当的统计、机器学习、深度学习等方法，对收集的大量数据进行计算、分析、汇总和整理，以求最大化地开发数据价值，发挥数据作用。

AI 开发的基本流程

AI开发的基本流程通常可以归纳为几个步骤：确定目的、准备数据、训练模型、评估模型、部署模型。

图 3-1 AI 开发流程



步骤1 确定目的

在开始AI开发之前，必须明确要分析什么？要解决什么问题？商业目的是什么？基于商业的理解，整理AI开发框架和思路。例如，图像分类、物体检测等等。不同的项目对数据的要求，使用的AI开发手段也是不一样的。

步骤2 准备数据

数据准备主要是指收集和预处理数据的过程。

按照确定的分析目的，有目的性的收集、整合相关数据，数据准备是AI开发的一个基础。此时最重要的是保证获取数据的真实可靠性。而事实上，不能一次性将所有数据

都采集全，因此，在数据标注阶段你可能会发现还缺少某一部分数据源，反复调整优化。

步骤3 训练模型

俗称“建模”，指通过分析手段、方法和技巧对准备好的数据进行探索分析，从中发现因果关系、内部联系和业务规律，为商业目的提供决策参考。训练模型的结果通常是一个或多个机器学习或深度学习模型，模型可以应用到新的数据中，得到预测、评价等结果。

业界主流的AI引擎有TensorFlow、Spark_MLlib、MXNet、Caffe、PyTorch、XGBoost-Sklearn、MindSpore等，大量的开发者基于主流AI引擎，开发并训练其业务所需的模型。

步骤4 评估模型

训练得到模型之后，整个开发过程还不算结束，需要对模型进行评估和考察。往往不能一次性获得一个满意的模型，需要反复的调整算法参数、数据，不断评估训练生成的模型。

一些常用的指标，如准确率、召回率、AUC等，能帮助您有效的评估，最终获得一个满意的模型。

步骤5 部署模型

模型的开发训练，是基于之前的已有数据（有可能是测试数据），而在得到一个满意的模型之后，需要将其应用到正式的实际数据或新产生数据中，进行预测、评价、或以可视化和报表的形式把数据中的高价值信息以精辟易懂的形式提供给决策人员，帮助其制定更加正确的商业策略。

----结束

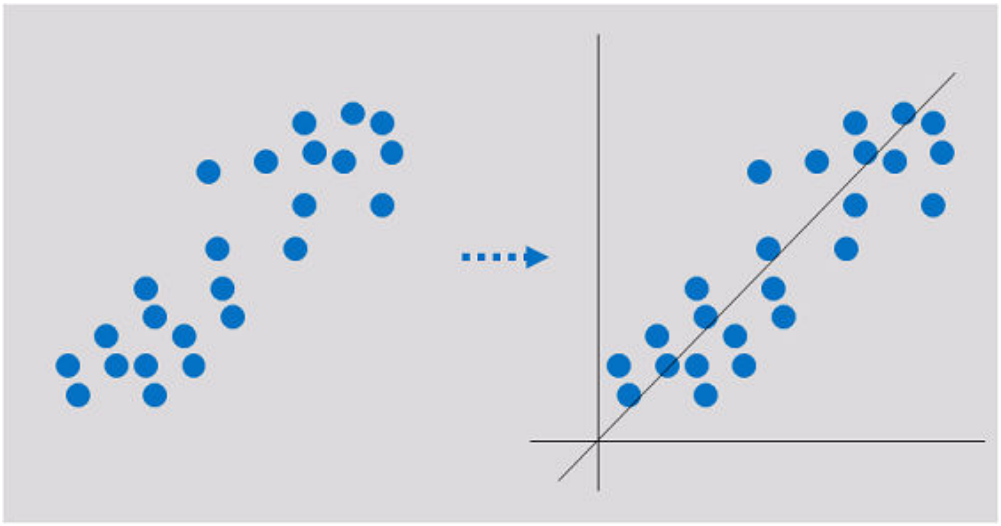
3.2 AI 开发基本概念

机器学习常见的分类有3种：

- 监督学习：利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程，也称为监督训练或有教师学习。常见的有回归和分类。
- 非监督学习：在未加标签的数据中，试图找到隐藏的结构。常见的有聚类。
- 强化学习：智能系统从环境到行为映射的学习，以使奖励信号（强化信号）函数值最大。

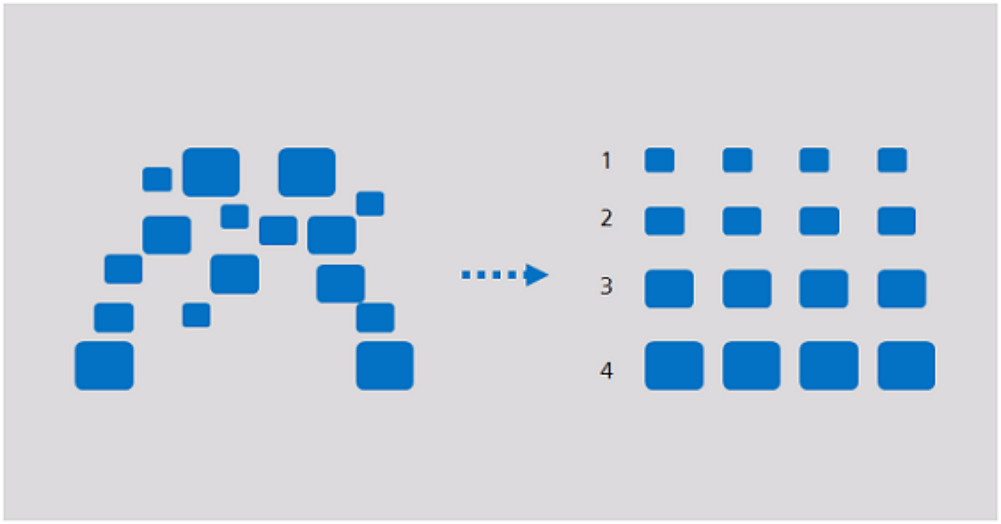
回归

回归反映的是数据属性值在时间上的特征，产生一个将数据项映射到一个实值预测变量的函数，发现变量或属性间的依赖关系，其主要研究问题包括数据序列的趋势特征、数据序列的预测以及数据间的关系等。它可以应用到市场营销的各个方面，如客户寻求、保持和预防客户流失活动、产品生命周期分析、销售趋势预测及有针对性的促销活动等。



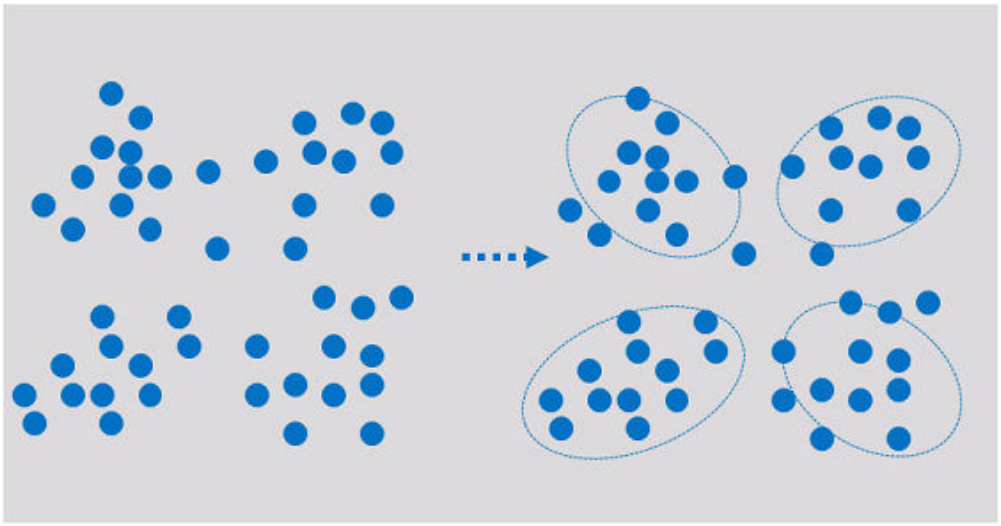
分类

分类是找出一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据项映射到某个给定的类别。它可以应用到客户的分类、客户的属性和特征分析、客户满意度分析、客户的购买趋势预测等。



聚类

聚类是把一组数据按照相似性和差异性分为几个类别，其目的是使得属于同一类别的数据间的相似性尽可能大，不同类别中的数据间的相似性尽可能小。它可以应用到客户群体的分类、客户背景分析、客户购买趋势预测、市场的细分等。



与分类不同，聚类分析数据对象，而不考虑已知的类标号（一般训练数据中不提供类标号）。聚类可以产生这种标号。对象根据最大化类内的相似性、最小化类间的相似性的原则进行聚类或分组。对象的聚类是这样形成的，使得在一个聚类中的对象具有很高的相似性，而与其它聚类中的对象很不相似。

3.3 ModelArts 中常用概念

自动学习

自动学习功能可以根据标注数据自动设计模型、自动调参、自动训练、自动压缩和部署模型，不需要代码编写和模型开发经验。只需三步，标注数据、自动训练、部署模型，即可完成模型构建。

端-边-云

端-边-云分别指端侧设备、智能边缘设备、公有云。

推理

指按某种策略由已知判断推出新判断的思维过程。人工智能领域下，由机器模拟人类智能，使用构建的神经网络完成推理过程。

在线推理

在线推理是对每一个推理请求同步给出推理结果的在线服务（Web Service）。

批量推理

批量推理是对批量数据进行推理的批量作业。

Ascend 芯片

Ascend芯片是华为设计的高计算力低功耗的AI芯片。

资源池

ModelArts提供的大规模计算集群，可应用于模型开发、训练和部署。支持公共资源池和专属资源池两种，分别为共享资源池和独享资源池。ModelArts默认提供公共资源池，按需计费。专属资源池需单独购买，专属使用，不与其他用户共享。

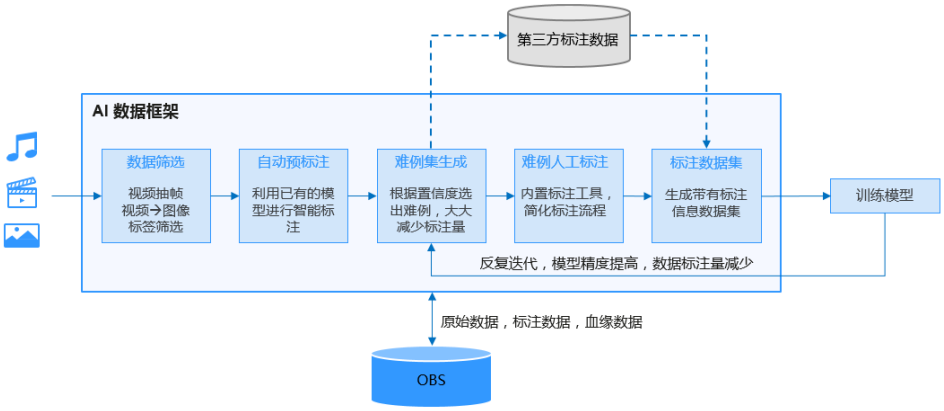
AI 市场

预置常用模型和算法，您可以直接获取使用。您也可以将自己开发的模型或算法分享至市场，共享给个人或者公开共享。

3.4 数据管理

AI 开发过程中经常需要处理海量数据，数据准备与标注往往耗费整体开发一半以上时间。ModelArts数据管理提供了一套高效便捷的管理和标注数据框架。不仅支持图片、文本、语音、视频等多种数据类型，涵盖图像分类、目标检测、音频分割、文本分类等多个标注场景，可适用于各种AI项目，如计算机视觉、自然语言处理、音视频分析等；同时提供数据筛选、数据分析、数据处理、智能标注、团队标注以及版本管理等功能，AI开发者可基于该框架实现数据标注全流程处理。如图所示。

图 3-2 数据标注全流程



ModelArts数据管理为数据集提供聚类分析、数据清洗、数据增强、数据选择、特征分析等处理，可帮助开发者进一步理解数据、筛选数据、挖掘数据信息，从而准备出一份满足开发目标或项目要求的高价值数据。

开发者可利用数据管理提供的各个场景标注工具进行数据标注，也可以选择多种标注方式，包括通过预置算法或用户自定义算法训练得到的模型进行智能标注，仅需少量人工标注和修正则可以得到较准确的标注结果；通过创建团队进行多人合作标注，提升标注效率。满足个人开发者的独立标注、小团队的协作标注，和专业团队的大规模协同标注及项目化管理。

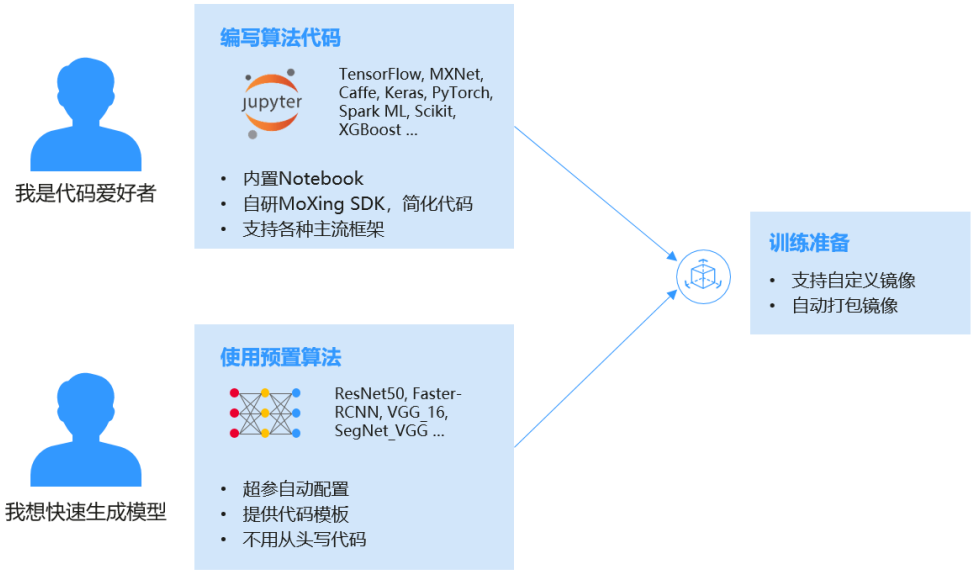
针对大规模团队的标注，提供专业的团队管理、人员管理、数据管理，实现从项目创建、分配、管理、标注、验收全流程。针对个人、小团队、小规模协作标注，提供便捷易用的标注工具，最小化项目管理开销。

此外，标注平台确保用户数据安全性，确保用户数据仅在授权范围内使用，标注对象分配策略确保用户数据的隐私性，实现标注数据脱敏需求。

3.5 开发环境

在 AI 开发过程中搭建开发环境、选择AI算法框架、选择算法、调试代码、安装相应软件或者硬件加速驱动库都不是容易的事情，使得学习 AI 开发上手慢门槛高。为了解决这些问题，ModelArts算法开发平台简化了整个开发过程，以降低开发门槛，算法开发过程如图3-3所示。

图 3-3 算法开发



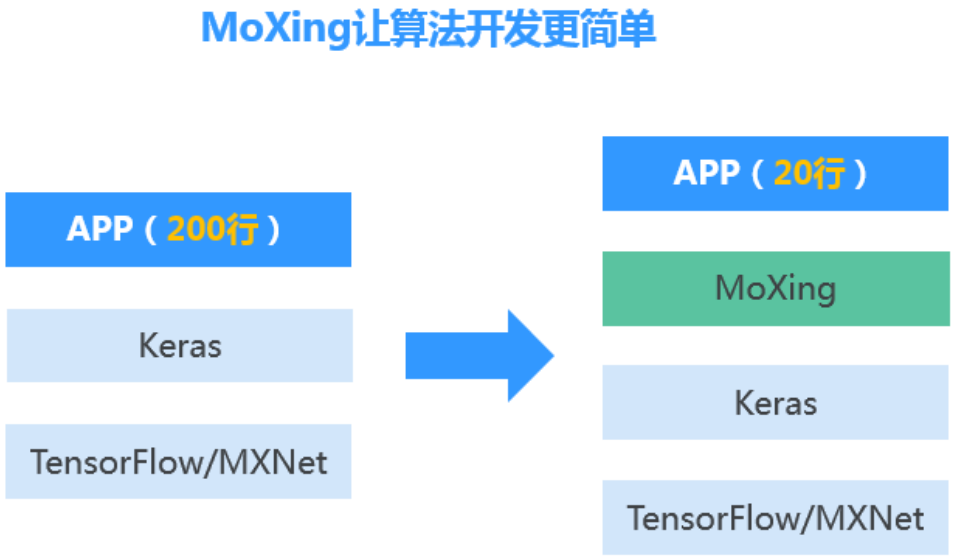
- **支持所有主流的 AI 算法框架**

  机器学习和深度学习领域，主流开源的训练和推理计算框架包括TensorFlow、PyTorch、MXNet、MindSpore等。ModelArts平台为适应不同开发者的开发习惯及不同应用场景，支持所有主流AI计算框架，并提供友好易用的开发和调测环境。支持传统机器学习算法运行，如逻辑回归、决策树、聚类算法等；支持CNN、RNN、LSTM等多种类型的深度学习算法执行。

- **简化面向分布式训练的算法开发**

  深度学习需要大规模的加速计算，往往需要大规模GPU集群进行分布式加速。而现有的开源框架需要算法开发者写大量的代码实现在不同硬件上的分布式训练，而且不同框架的加速代码都不相同。为了解决这些痛点，需要一种轻型的分布式框架或者SDK，构建于TensorFlow、PyTorch、MXNet、MindSpore等深度学习引擎之上，使得这些计算引擎分布式性能更高，同时易用性更好，ModelArts的MoXing可以很好地解决这个痛点。开发者基于MoXing开发的代码如图3-4所示。

图 3-4 基于 MoXing 开发



- 简化调参，集成多种调参技巧包，如数据增强的调参策略，可简化AI算法工程师的模型调优痛苦。
- 简化分布式，支持将单机代码自动分布式，使算法工程师不需要学习分布式相关的知识，在自动化分布式的同时，也优化了分布式的性能，自动化和高性能是相辅相成的。

3.6 模型训练

模型训练中除了数据和算法外，开发者花了大量时间在模型参数设计上。模型训练的参数直接影响模型的精度以及模型收敛时间，参数的选择极大依赖于开发者的经验，参数选择不当会导致模型精度无法达到预期结果，或者模型训练时间大大增加。

为了降低开发者的专业要求，提升开发者模型训练的开发效率及训练性能，ModelArts基于机器学习算法及强化学习的模型训练自动超参调优，如learning rate、batch size等自动的调参策略；预置和调优常用模型，简化模型开发。

当前大多数开发者开发模型时，为了满足精度需求，模型往往达到几十层，甚至上百层，参数规模达到百兆甚至在GB规格以上，导致对计算资源的规格要求极高，主要体现在对硬件资源的算力及内存、ROM的规格的需求上。端侧资源规格限制极为严格，以端侧智能摄像头为例，通常端侧算力在1TFLOPS，内存在2GB规格左右，ROM空间在2GB左右，需要将端侧模型大小控制在百KB级别，推理时延控制在百毫秒级别。

这就需要借助模型精度无损或微损下的压缩技术，如通过剪枝、量化、知识蒸馏等技术，实现模型的自动压缩及调优，进行模型压缩和重新训练的自动迭代，以保证模型的精度损失极小。无需重新训练的低比特量化技术实现模型从高精度浮点向定点运算转换，多种压缩技术和调优技术实现模型计算量满足端、边小硬件资源下的轻量化需求，模型压缩技术在特定领域场景下实现精度损失<1%。

当训练数据量很大时，深度学习模型的训练将会非常耗时。在计算机视觉中，ImageNet-1k（包含 1000 个类别的图像分类数据集，以下简称 ImageNet）是经典、常用的一个数据集，如果我们在该数据集上用一块P100 GPU训练一个ResNet-50模型，则需要耗时将近1周，严重阻碍了深度学习应用的开发进度。因此，深度学习训练加速一直是学术界和工业界所关注的重要问题。

分布式训练加速需要从软硬件两方面协同来考虑，仅单一的调优手段无法达到期望的加速效果。所以分布式加速的调优是一个系统工程，需要从硬件角度（芯片、硬件设计）考虑分布式训练架构，如系统的整体计算规格、网络带宽、高速缓存、功耗、散热等因素，充分考虑计算和通信的吞吐量关系，以实现计算和通信时延的隐藏。

软件设计需要结合高性能硬件特性，充分利用硬件高速网络实现高带宽分布式通信，实现高效的数据集本地数据缓存技术，通过训练调优算法，如混合并行，梯度压缩、卷积加速等技术，实现分布式训练系统软硬件端到端的高效协同优化，实现多机多卡分布式环境下训练加速。ModelArts 在千级别资源规格多机多卡分布式环境下，典型模型 ResNet50 在 ImageNet 数据集上实现加速比>0.8，是行业领先水平。

衡量分布式深度学习的加速性能时，主要有如下2个重要指标：

- 吞吐量，即单位时间内处理的数据量。
- 收敛时间，即达到一定的收敛精度所需的时间。

吞吐量一般取决于服务器硬件（如更多、更大FLOPS处理能力的AI加速芯片，更大的通信带宽等）、数据读取和缓存、数据预处理、模型计算（如卷积算法选择等）、通信拓扑等方面的优化。除了低bit计算和梯度（或参数）压缩等，大部分技术在提升吞吐量的同时，不会造成对模型精度的影响。为了达到最短的收敛时间，需要在优化吞吐量的同时，对调参方面也做调优。调参不到位会导致吞吐量难以优化，当batch size超参不足够大时，模型训练的并行度就会相对较差，吞吐量难以通过增加计算节点个数而提升。

对用户而言，最终关心的指标是收敛时间，因此ModelArts的MoXing实现了全栈优化，极大缩短了训练收敛时间。在数据读取和预处理方面，MoXing通过利用多级并发输入流水线使得数据 IO 不会成为瓶颈；在模型计算方面，MoXing对上层模型提供半精度和单精度组成的混合精度计算，通过自适应的尺度缩放减小由于精度计算带来的损失；在超参调优方面，采用动态超参策略（如 momentum、batch size等）使得模型收敛所需epoch个数降到最低；在底层优化方面，MoXing与底层华为服务器和通信计算库相结合，使得分布式加速进一步提升。

ModelArts 高性能分布式训练优化点

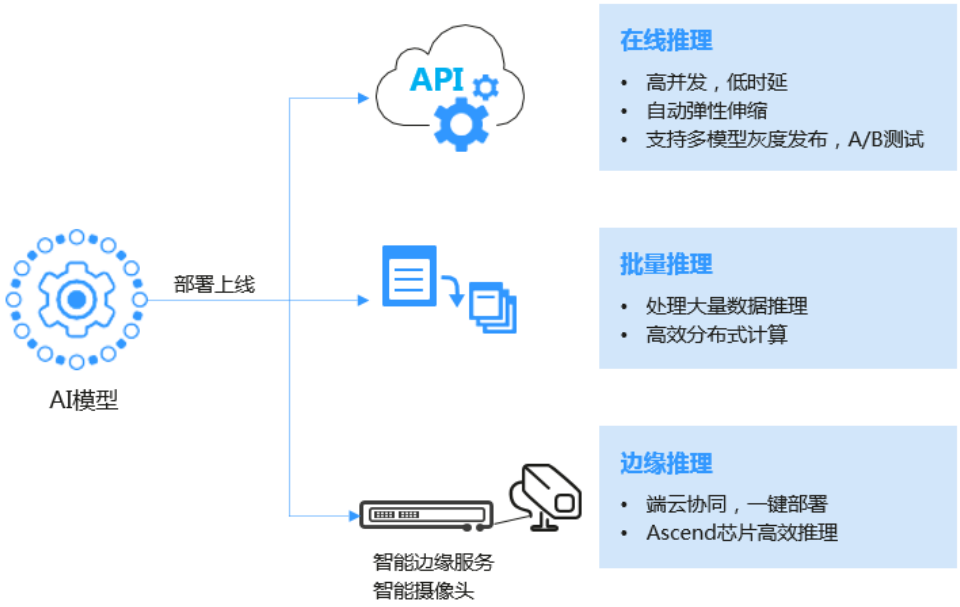
- 自动混合精度训练（充分发挥硬件计算能力）
- 动态超参调整技术（动态 batch size、image size、momentum 等）
- 模型梯度的自动融合、拆分
- 基于BP bubble自适应的计算，通信算子调度优化
- 分布式高性能通信库（nstack、HCCL）
- 分布式数据-模型混合并行
- 训练数据压缩、多级缓存

3.7 模型部署

通常AI模型部署和规模化落地非常复杂。

例如，智慧交通项目中，在获得训练好的模型后，需要部署到云、边、端多种场景。如果在端侧部署，需要一次性部署到不同规格、不同厂商的摄像机上，这是一项非常耗时、费力的巨大工程，ModelArts支持将训练好的模型一键部署到端、边、云的各种设备上和各种场景上，并且还为个人开发者、企业和设备生产厂商提供了一整套安全可靠的一站式部署方式。

图 3-5 部署模型的流程



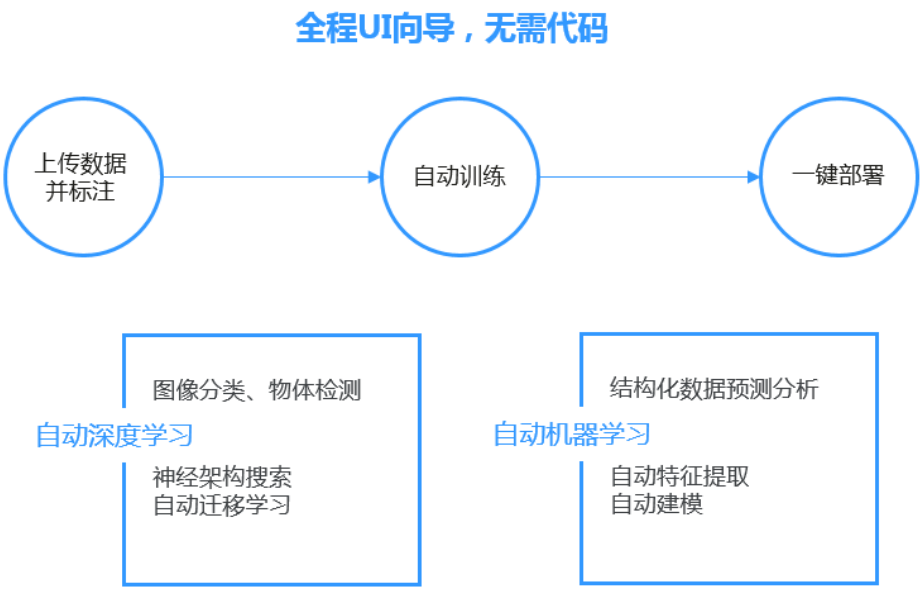
- 在线推理服务，可以实现高并发，低延时，弹性伸缩，并且支持多模型灰度发布、A/B测试。
- 支持各种部署场景，既能部署为云端的在线推理服务和批量推理任务，也能部署到端，边等各种设备。
- 一键部署，可以直接推送部署到边缘设备中，选择智能边缘节点，推送模型。
- ModelArts基于Ascend 310高性能AI推理芯片的深度优化，具有PB级别的单日推理数据处理能力，支持发布云上推理的API百万个以上，推理网络时延毫秒。

3.8 自动学习

AI 要规模化走进各行各业，必须要降低AI模型开发难度和门槛。当前仅少数算法工程师和研究员掌握AI的开发和调优能力，并且大多数算法工程师仅掌握算法原型开发能力，缺少相关的原型到真正产品化、工程化的能力。而对于大多数业务开发者来说，更是不具备AI算法的开发和参数调优能力。这导致大多数企业都不具备AI开发能力。

ModelArts通过机器学习的方式帮助不具备算法开发能力的业务开发者实现算法的开发，基于迁移学习、自动神经网络架构搜索实现模型自动生成，通过算法实现模型训练的参数自动化选择和模型自动调优的自动学习功能，让零AI基础的业务开发者可快速完成模型的训练和部署。依据开发者提供的标注数据及选择的场景，无需任何代码开发，自动生成满足用户精度要求的模型。可支持图片分类、物体检测、预测分析、声音分类场景。可根据最终部署环境和开发者需求的推理速度，自动调优并生成满足要求的模型。

图 3-6 自动学习流程



ModelArts 的自动学习不止为入门级开发者使用设计，还提供了“自动学习白盒化”的能力，开放模型参数，实现模板化开发。很多资深的开发者说，希望有一款工具，可以自动生成模型，然后在这个基础上修改，这很像普通软件的模板化开发，在一个半成品的基础上调优，重新训练模型，提高开发效率。

自动学习的关键技术主要是基于信息熵上限近似模型的树搜索最优特征变换和基于信息熵上限近似模型的贝叶斯优化自动调参。通过这些关键技术，可以从企业关系型（结构化）数据中，自动学习数据特征和规律，智能寻优特征&ML模型及参数，准确性达到甚至专家开发者的调优水平。自动深度学习的关键技术主要是迁移学习（只通过少量数据生成高质量的模型），多维度下的模型架构自动设计（神经网络搜索和自适应模型调优），和更快、更准的训练参数自动调优自动训练。

4 ModelArts 支持的 AI 框架

ModelArts的开发环境、训练作业、模型推理（即模型管理和部署上线）支持的AI框架及其版本，不同模块的呈现方式存在细微差异，各模块支持的AI框架请参见如下描述。

开发环境

开发环境的Notebook，根据不同的工作环境（即不同Python版本），对应支持的AI引擎和版本有所不同，在创建好对应工作环境的Notebook实例后，再根据表4-1对应的版本创建文件。而且Notebook支持多引擎，即同一个Notebook实例，可以使用所有支持的AI引擎，不同引擎之间可快速、方便切换。

表 4-1 AI 引擎

工作环境名称	预置的AI引擎及版本	适配芯片
Multi-Engine 1.0 (python3 推荐)	MXNet-1.2.1	CPU/GPU
	PySpark-2.3.2	CPU
	Pytorch-1.0.0	GPU
	TensorFlow-1.13.1	CPU/GPU
	TensorFlow-1.8	CPU/GPU
	XGBoost-Sklearn	CPU
Multi-Engine 1.0 (python2)	Caffe-1.0.0	CPU/GPU
	MXNet-1.2.1	CPU/GPU
	PySpark-2.3.2	CPU
	PyTorch1.0.0	GPU
	TensorFlow-1.13.1	CPU/GPU
	TensorFlow-1.8	CPU/GPU
	XGBoost-Sklearn	CPU

工作环境名称	预置的AI引擎及版本	适配芯片
Multi-Engine 2.0 (python3)	Pytorch-1.4.0	GPU
	R-3.6.1	CPU/GPU
	TensorFlow-2.1.0	CPU/GPU

训练作业

创建训练作业时，支持的AI引擎及对应版本如下所示。

- **TensorFlow**：TF-1.8.0-python3.6、TF-1.8.0-python2.7、TF-1.13.1-python3.6、TF-1.13.1-python2.7、TF-2.1.0-python3.6
- **MXNet**：MXNet-1.2.1-python3.6、MXNet-1.2.1-python2.7
- **Caffe**：Caffe-1.0.0-python2.7
- **Spark_MLlib**：Spark-2.3.2-python2.7、Spark-2.3.2-python3.6
- **Ray**：RAY-0.7.4-python3.6
- **XGBoost-Sklearn**：XGBoost-0.80-Sklearn-0.18.1-python2.7、XGBoost-0.80-Sklearn-0.18.1-python3.6
- **PyTorch**：PyTorch-1.0.0-python2.7、PyTorch-1.0.0-python3.6、PyTorch-1.3.0-python2.7、PyTorch-1.3.0-python3.6、PyTorch-1.4.0-python3.6
- **Ascend-Powered-Engine**：MindSpore-1.0-python3.7-aarch64、TF-1.15-python3.7-aarch64

说明

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

模型推理

针对导入模型，并在ModelArts完成模型推理的。支持如下常用引擎及版本。

表 4-2 支持的常用引擎及其 Runtime

模型使用的引擎类型	支持的运行环境（Runtime）	注意事项
TensorFlow	python3.6 python2.7 tf1.13-python2.7-gpu tf1.13-python2.7-cpu tf1.13-python3.6-gpu tf1.13-python3.6-cpu tf1.13-python3.7-cpu tf1.13-python3.7-gpu tf2.1-python3.7	• python2.7、python3.6、python3.7的运行环境搭载的TensorFlow版本为1.8.0。 • python3.6、python2.7、tf2.1-python3.7，表示该模型可同时在CPU或GPU运行。其他Runtime的值，如果后缀带cpu或gpu，表示该模型仅支持在CPU或GPU中运行。 • 默认使用的Runtime为python2.7。

模型使用的引擎类型	支持的运行环境（Runtime）	注意事项
MXNet	python3.7 python3.6 python2.7	● python2.7、python3.6、python3.7的运行环境搭载的MXNet版本为1.2.1。 ● python2.7、python3.6、python3.7，表示该模型可同时在CPU或GPU运行。 ● 默认使用的Runtime为python2.7。
Caffe	python2.7 python3.6 python3.7 python2.7-gpu python3.6-gpu python3.7-gpu python2.7-cpu python3.6-cpu python3.7-cpu	● python2.7、python3.6、python3.7、python2.7-gpu、python3.6-gpu、python3.7-gpu、python2.7-cpu、python3.6-cpu、python3.7-cpu的运行环境搭载的Caffe版本为1.0.0。 ● python2.7、python3.6、python3.7只能用于运行适用于CPU的模型。其他Runtime的值，如果后缀带cpu或gpu，表示该模型仅支持在CPU或GPU中运行。推荐使用python2.7-gpu、python3.6-gpu、python3.7-gpu、python2.7-cpu、python3.6-cpu、python3.7-cpu的Runtime。 ● 默认使用的Runtime为python2.7。
Spark_MLlib	python2.7 python3.6	● python2.7以及python3.6的运行环境搭载的Spark_MLlib版本为2.3.2。 ● 默认使用的Runtime为python2.7。 ● python2.7、python3.6只能用于运行适用于CPU的模型。
Scikit_Learn	python2.7 python3.6	● python2.7以及python3.6的运行环境搭载的Scikit_Learn版本为0.18.1。 ● 默认使用的Runtime为python2.7。 ● python2.7、python3.6只能用于运行适用于CPU的模型。
XGBoost	python2.7 python3.6	● python2.7以及python3.6的运行环境搭载的XGBoost版本为0.80。 ● 默认使用的Runtime为python2.7。 ● python2.7、python3.6只能用于运行适用于CPU的模型。

模型使用的引擎类型	支持的运行环境（Runtime）	注意事项
PyTorch	python2.7 python3.6 python3.7 pytorch1.4-python3.7	• python2.7、python3.6、python3.7的运行环境搭载的PyTorch版本为1.0。 • python2.7、python3.6、python3.7、pytorch1.4-python3.7，表示该模型可同时在CPU或GPU运行。 • 默认使用的Runtime为python2.7。
MindSpore	python3.7	python3.7的运行环境搭载的MindSpore版本为1.0。

5 与其他服务的关系

与统一身份认证服务的关系

ModelArts使用统一身份认证服务（Identity and Access Management，简称IAM）实现认证功能。IAM的更多信息请参见**《统一身份认证服务用户指南》**。

与对象存储服务的关系

ModelArts使用对象存储服务（Object Storage Service，简称OBS）存储数据和模型的备份和快照，实现安全、高可靠和低成本的存储需求。OBS的更多信息请参见**《对象存储服务控制台指南》**。

表 5-1 ModelArts 各环节与 OBS 的关系

功能	子任务	ModelArts与OBS的关系
自动学习	数据标注	ModelArts标注的数据存储在OBS中。
	自动训练	训练作业结束后，其生成的模型存储在OBS中。
	部署模型	ModelArts将存储在OBS中的模型部署上线为在线服务。
AI全流程开发	数据管理	• 数据集存储在OBS中。 • 数据集的标注信息存储在OBS中。 • 支持从OBS中导入数据。
	开发环境	Notebook实例中的数据或代码文件存储在OBS中。
	训练模型	• 训练作业使用的数据集存储在OBS中。 • 训练作业的运行脚本存储在OBS中。 • 训练作业输出的模型存储在指定的OBS中。 • 训练作业的过程日志存储在指定的OBS中。
	模型管理	训练作业结束后，其生成的模型存储在OBS中，导入模型时，从OBS中导入已有的模型。

功能	子任务	ModelArts与OBS的关系
	部署上线	将存储在OBS中的模型部署上线。
全局配置	-	获取访问授权（使用委托或访问密钥授权），以便ModelArts可以使用OBS存储数据、创建Notebook等操作。

与云硬盘的关系

ModelArts使用云硬盘服务（Elastic Volume Service，简称EVS）存储创建的Notebook实例。EVS的更多信息请参见《**云硬盘用户指南**》。

与云容器引擎的关系

ModelArts使用云容器引擎（Cloud Container Engine，简称CCE）部署模型为在线服务，支持服务的高并发和弹性伸缩需求。CCE的更多信息请参见**《云容器引擎用户指南》**。

与云监控的关系

ModelArts使用云监控服务（Cloud Eye Service， 简称CES）监控在线服务和对应模型负载，执行自动实时监控、告警和通知操作。CES的更多信息请参见**《云监控服务用户指南》**。

与云审计的关系

ModelArts使用云审计服务（Cloud Trace Service，简称CTS）记录ModelArts相关的操作事件，便于日后的查询、审计和回溯。CTS的更多信息请参见**《云审计服务指南》**。

6 如何访问 ModelArts

云服务平台提供了Web化的服务管理平台，即管理控制台和基于HTTPS请求的API（Application programming interface）管理方式。

- **管理控制台方式**

  ModelArts提供了简洁易用的管理控制台，包含自动学习、数据管理、开发环境、模型训练、模型管理、部署上线、AI市场等功能，您可以在管理控制台端到端完成您的AI开发。

  使用ModelArts管理控制台，需先注册华为云。如果您已注册华为云，可从主页选择“EI 企业智能 > AI服务>EI 基础平台 > AI开发平台ModelArts”直接登录管理控制台。如果未注册，请参见**注册账号**。

- **SDK方式**

  如果您需要将ModelArts集成到第三方系统，用于二次开发，可选择调用SDK方式完成目的。ModelArts的SDK是对ModelArts服务提供的REST API进行的Python封装，简化用户的开发工作。具体操作和SDK详细描述，请参见《**SDK参考**》。

  除此之外，在管理控制台的Notebook中编写代码时，也可直接调用ModelArts SDK。

- **API方式**

  如果您需要将ModelArts集成到第三方系统，用于二次开发，请使用API方式访问ModelArts，具体操作和API详细描述，请参见《**API参考**》。

7 计费说明

ModelArts是面向AI开发者的一站式开发平台，提供海量数据预处理及半自动化标注、大规模分布式训练、自动化模型生成及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期AI工作流。

ModelArts服务的计费方式简单、灵活，您既可以选择按实际使用时长计费。也可以选择更经济的按包周期计费方式。

计费项

ModelArts服务根据用户选择使用的资源不同进行收费。具体计费项请参见表7-1，每个计费项的详细价格请参见：**产品价格详情**。

表 7-1 计费项说明

计费项	说明
AI全流程开发	面向有AI基础的开发者，提供机器学习和深度学习的算法开发及部署全功能，包含数据处理、模型开发、模型训练、模型管理和部署上线流程。涉及计费项包含：模型开发环境（Notebook）、模型训练（训练作业、TensorBoard）、部署上线（在线服务）。
自动学习	面向AI基础能力弱的开发者，根据标注数据、自动设计、调优、训练模型和部署服务，根据开发者零编码实现模型定制化开发。此计费资源仅适用于自动学习作业的训练和部署。涉及计费项包括：自动学习（训练作业）、自动学习（部署上线）。 当前仅支持按需付费模式。

计费模式

ModelArts主要提供按需和预付套餐包的计费方式供您灵活选择。

- **按需购买**：这种购买方式比较灵活，可以即开即停。
- **预付套餐包**：客户预先购买一定的资源使用量配额，在按需使用过程中，系统优先扣减配额，超出配额的使用量才需要额外根据按需费用付费。

- **包周期（包年包月）购买**：华为云提供包年和包月的购买模式。这种购买方式相对于按需付费则能够提供更大的折扣。

  说明

  目前，只有“专属资源池”支持包周期购买模式。

变更配置

在使用ModelArts时，您可根据业务需要选择合适的资源。当作业启动后，您可以使用如下变更配置的方式。

- 当您目前购买的资源不满足业务需求时，请参见**购买套餐包**购买更高规格的配置资源。
- 专属资源池扩缩容：包周期（包年包月）的专属资源池不支持扩缩容。如果购买的是“按需计费”的专属资源池，那么您可以手动扩缩容，计费会按照修改后的节点数量进行收费，具体操作请参见**扩缩容专属资源池**。

若ModelArts提供的变更配置方式不满足您的要求，您也可以通过重建作业，做数据迁移的方式实现配置变更。

续费

目前ModelArts提供按需和预付套餐包购买方式，按需是每小时扣费，如果余额不足会导致欠费。而预付套餐包是超出当前套餐包的额度后系统会自动以按需计费的方式进行结算，只要您的账户上有足够余额，则不会影响您的使用，如果余额不足会导致欠费。如果您未能续费，华为云不会立即停止您的业务，订单转入保留期，此时将终止服务，数据仍然保留。

- 保留期的时长由客户等级而定，具体请参见**保留期**。
- 如需续费，请进入**续费管理**页面进行续费操作。

欠费与到期

- 按需计费模式和预付套餐包的资源，没有到期的概念。预付套餐包超出当前套餐包的额度将自动转为按需计费。按需购买的资源是按每小时扣费，当余额不足，无法对上一个小时的费用进行扣费，就会导致欠费，欠费后有**保留期**。您续费后解冻资源，可继续正常使用，请注意在保留期进行的续费，是以原到期时间作为生效时间，您应当支付从进入保留期开始到续费时的服务费用。

您购买的资源欠费后，会导致部分操作受限，建议您尽快续费。具体受限操作如**表7-2**所示：

表 7-2 欠费受限操作

功能	受限操作
自动学习	模型训练、部署上线
数据管理-数据集	一键模型上线任务
开发环境-Notebook	创建Notebook、启动Notebook
训练管理-训练作业	创建训练作业

功能	受限操作
训练管理-自动化搜索作业	创建自动化搜索作业
部署上线-在线服务	部署在线服务
专属资源池	创建专属资源池
AI市场-ModelHub	创建模型、算法或HiLens技能

8 权限管理

如果您需要对购买的ModelArts资源，给企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制资源的访问。

通过IAM，您可以在账号中给员工创建IAM用户，并使用策略来控制他们对资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有ModelArts的使用权限，但是不希望他们拥有删除ModelArts训练作业等高危操作的权限，那么您可以使用IAM为开发人员创建用户，通过授予仅能使用ModelArts，但是不允许删除ModelArts训练作业的权限策略，控制他们对ModelArts资源的使用范围。

如果账号已经能满足您的要求，不需要创建独立的IAM用户进行权限管理，您可以跳过本章节，不影响您使用ModelArts服务的其它功能。

IAM是提供权限管理的基础服务，无需付费即可使用，您只需要为您账号中的资源进行付费。关于IAM的详细介绍，请参见**《IAM产品介绍》**。

ModelArts 权限

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于授予的权限对云服务进行操作。

ModelArts部署时通过物理区域划分，为项目级服务，授权时，“作用范围”需要选择“区域级项目”，然后在指定区域（如华北-北京1）对应的项目（cn-north-1）中设置相关权限，并且该权限仅对此项目生效；如果在“所有项目”中设置权限，则该权限在所有区域项目中都生效。访问ModelArts时，需要先切换至授权区域。

根据授权精细程度分为角色和策略。

- 策略：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。例如：针对ECS服务，管理员能够控制IAM用户仅能对某一类云服务器资源进行指定的管理操作。ModelArts支持的API授权项请参见**《API参考》>权限策略和授权项**。

如**表8-1**所示，包括了ModelArts的所有系统权限。

表 8-1 ModelArts 系统策略

策略名称	描述	策略类别
ModelArts FullAccess	ModelArts管理员用户，拥有所有ModelArts服务的权限	系统策略
ModelArts CommonOperations	ModelArts操作用户，拥有所有ModelArts服务操作权限除了管理专属资源池的权限	系统策略

说明

当为IAM用户配置ModelArts权限时，需同时为其配置对应的OBS服务权限，才可以正常使用OBS的各项功能。

- 当您需要为用户授予OBS的管理员操作权限时，需为IAM用户配置“作用范围”为“全局级服务”的“Tenant Administrator”策略，详细说明请参见**OBS权限管理**。
- 当您需要限制用户操作，仅为ModelArts用户配置OBS相关的最小化权限项，具体操作请参见**创建ModelArts自定义策略**。

表8-2列出了ModelArts常用操作与系统策略的授权关系，您可以参照该表选择合适的系统策略。

表 8-2 常用操作与系统策略的授权关系

操作	ModelArts FullAccess	ModelArts CommonOperations
自动学习	√	√
数据标注	√	√
数据管理	√	√
开发环境	√	√
模型管理	√	√
部署上线	√	√
AI市场	√	√
专属资源池	√	x
全局配置	√	√

相关链接

- **IAM产品介绍**
- **创建用户组、用户并授予ModelArts权限**
- **策略支持的授权项**

9 配额说明

本服务应用的基础设施如下：

- 弹性云服务器
- 云硬盘
- 虚拟私有云

其配额查看及修改请参见**关于配额**。

