

TI-ONE 训练平台

快速入门



腾讯云

【版权声明】

©2013–2025 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【商标声明】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【服务声明】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。

您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

文档目录

快速入门

- 平台使用概述

- 平台使用准备

 - 配置账号和权限

 - 开通关联产品

- 腾讯混元大模型系列

 - Hunyuan-Large x TI 上手指南

 - 精调 Hunyuan-Large 实践教程

- LLM 大模型系列

 - 快速部署和体验内置开源 LLM 大模型

 - 导入和部署自定义 LLM 大模型（平台内置推理镜像）

 - 精调内置开源大模型

- 传统 AI 模型系列

 - 使用任务式建模构建手写体分类模型

 - 导入和部署第三方 AI 模型

- Tikit 快速入门

快速入门

平台使用概述

最近更新时间：2025-01-17 17:32:12

总览

本文介绍 TI-ONE 训练平台使用前置准备工作，根据不同背景的用户，提供不同的文档和流程指引，介绍平台的完整使用流程。

TI-ONE 使用前置准备工作

TI-ONE 使用前需要先 [开通关联产品](#)。
使用按量计费或者在资源组中 [购买资源](#)。

不同背景用户使用指引

若您希望体验平台内置的大模型，并进行推理验证。可参考 [快速试一试开源大模型](#)。

若您是有 AI 开发经验、熟悉代码编写的用户，您可以使用 Notebook 开发环境，开发训练代码进行模型开发和调试。您可以参考 [使用 Notebook 完成图像分类模型构建](#)。

如果您已经有模型镜像，希望接入到 TI-ONE 平台进行训练和推理，您可以参考文档 [自定义训练镜像规范](#) 完成镜像改造后，在任务式建模中选择自定义镜像方式进行训练，[使用自定义镜像发布在线服务开发指引](#)。

TI-ONE 完整使用流程

机器学习开发的核心流程通常可以分几个步骤：数据准备、模型开发、模型训练、模型管理、模型部署。接下来介绍如何在TI-ONE 平台，从数据准备到完成模型开发和部署的全流程。

步骤	任务	说明	详细指导
数据准备	导入数据集	基于您准备的数据，可以在 TI-ONE 中上传、导入数据集。	导入数据集
	数据构建	帮助您从一份原始数据构建出高质量的 LLM 训练数据。	使用数据构建准备 LLM 精调数据
	数据标注	根据您创建的数据集，标注数据，方便后续训练使用。	数据标注
模型开发	新建 Notebook	新建一个 Notebook 作为开发环境，编写代码调试模型。	新建Notebook实例
模型训练	创建训练任务	创建一个任务式建模任务，选择数据集存储，并挂载前面编写完成的训练代码。训练完成后，将生成模型保存至输出的存储中。也支持一键发布模型至模型仓库。	创建任务式建模任务
模型管理	导入模型	将训练输出的模型导入至模型仓库，方便管理，支持将模型部署为在线服务。	模型包规范
模型部署	部署服务	将模型部署为在线服务。	部署在线服务
	访问服务	在线服务部署完成后，您可以访问测试。支持公网、VPC等多种在线服务调用方式。	在线服务调用

平台使用准备

配置账号和权限

最近更新时间：2025-05-14 17:03:52

总览

TI-ONE 训练平台使用过程需要使用其他云产品的API（例如：COS、CFS、GooseFSx 等），因此在正式使用前，需要提前开通对应访问授权。本文档介绍主账号、子账号和对应支持的权限策略。

主账号授权

前提条件

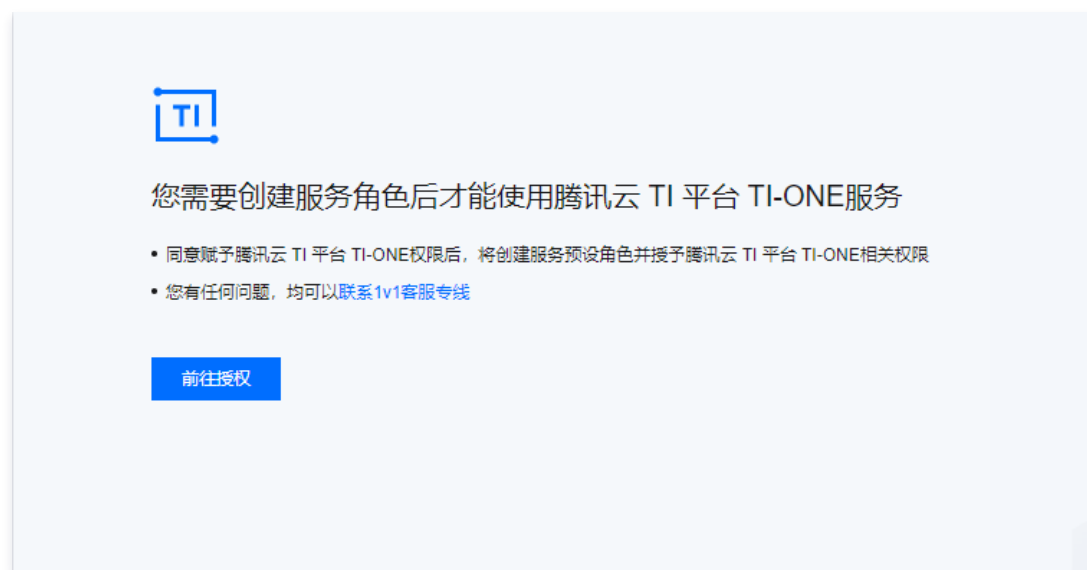
[注册腾讯云账号](#)

说明

当您注册腾讯云账号后，系统默认为您创建了一个主账号，用于快捷访问腾讯云资源。

操作步骤

1. 使用 **主账号**登录 [TI-ONE 控制台](#)，提示需要创建服务角色权限以正常访问其他云产品资源，使 TI-ONE 正常运行。



2. 单击[前往授权](#)，进入 CAM 控制台授权，单击[同意授权](#)，则为 TI-ONE 平台授权服务角色访问您其他云产品资源。

← 角色管理

服务授权

同意赋予 腾讯云 TI 平台 TI-ONE 权限后，将创建服务预设角色并授予 腾讯云 TI 平台 TI-ONE 相关权限

角色名称

TIONE_QcsRole

角色类型

服务角色

角色描述

当前角色为 腾讯云 TI 平台 TI-ONE 服务角色，该角色将在已关联策略的权限范围内访问您的其他云服务资源。

授权策略

预设策略 QcloudTioneFullAccessContainMultiservice、预设策略 QcloudTioneResourceGroupFullAccessContainMultiservice、预设策略 QcloudTioneReadOnlyAccessContainMultiservice

同意授权

取消

子账号授权

1. TI-ONE 支持子账号、协作者账号登录，主账号可以授权子账号或协作者访问管理权限。
2. 为了满足主账号便捷给予子账号授予常用范围的权限，腾讯云 TI 平台预设了如下三个预设策略，主账号可根据对子账号的权限定位在CAM控制台进行授权。

预设策略名称	功能描述	常用场景
QcloudTioneFullAccessContainMultiservice	TI-ONE 平台全读写访问权限，以及其他云产品包括CAM、COS、CFS、VPC、监控、标签等和下单交易权限。 拥有该策略的子账号/协作者可以完整使用TI-ONE平台的所有功能	管理员
QcloudTioneResourceGroupFullAccessContainMultiservice	TI-ONE 资源组管理模块的读写权限，平台其他模块的只读权限，部分云产品包括CAM、COS、CFS、VPC、监控、标签等的只读权限，和下单交易权限。 拥有该策略的子账号/协作者有 TI-ONE 平台的只读权限和资源组管理模块的全读写权限	资源管理员
QcloudTioneReadOnlyAccessContainMultiservice	TI-ONE 平台只读权限，以及其他云产品包括CAM、COS、CFS、VPC、监控、标签等的只读权限。 拥有该策略的子账号/协作者拥有 TI-ONE 平台的只读权限	全平台只读

- 具体操作详见[授权管理](#)。
3. 当 2 中的预设策略无法满足主账号对子账号的自定义权限管控需求时，可以使用自定义策略为子账号授权，相关操作指引请查看 [CAM 策略授权使用说明](#)，详细接口说明请查看 [CAM 业务接口说明](#)。
4. 如果您已对资源绑定了标签，并希望给予子账号控制标签属性资源的访问权限，您需要通过 [按标签授权](#) 创建自定义策略。详情可参考 [使用 CAM 标签实现子用户间的 TI-ONE 资源隔离](#)。

开通关联产品

最近更新时间：2025-05-14 10:17:12

概览

使用 TI-ONE 训练平台过程中，不同的场景下会使用其他腾讯云产品，如 COS、CFS、CLS、TCR 等。因此在正式使用前，需要根据场景需要提前开通对应的其他云产品并做好授权，保障后续的使用顺利进行。本文介绍各场景下依赖的其他云产品列表及权限要求。

前置要求

- 1. 已注册完成腾讯云账号
- 2. 已完成实名认证，详细操作指引请参见[实名认证介绍](#)。

TI-ONE 产品关联云产品

主要产品功能	关联云产品	和 TI-ONE 的关系	开通指引
全平台	访问管理-CAM	账号权限控制	访问管理指引文档
数据中心	对象存储-COS	存储数据集和标注文件，支持从 COS 中导入数据集	对象存储指引文档
	CFS	支持新建数据集关联到已有的 CFS 文件路径	文件存储指引文档
训练工坊	对象存储-COS	任务式建模中的代码包、训练输出存储存储模型，代码，训练数据	对象存储指引文档
	CFS/TurboCFS	任务式建模的数据集 Notebook 实例中的数据或代码文件存储	文件存储指引文档
	GooseFS/GooseFSx	任务式建模的数据集	数据加速指引文档
	镜像服务	上传自定义镜像用于训练	容器镜像服务企业版指引文档 容器镜像服务个人版指引文档
	日志服务-CLS	投递日志用于日志分析，监控告警	日志服务指引文档
	腾讯云可观测平台-TCOP	训练任务资源使用监控告警	腾讯云可观测平台指引文档
模型仓库	对象存储-COS	模型存储，支持从 COS 中导入模型	对象存储指引文档
模型服务	CFS/TurboCFS	在线服务，将存储在 CFS 中的模型部署为在线服务	文件存储指引文档
	GooseFSx	在线服务，将存储在 GooseFSx 中的模型部署为在线服务	数据加速器 GooseFS 指引文档
	对象存储-COS	批量预测，数据从 COS 输入、输出	对象存储指引文档
	镜像服务	在线服务，上传自定义模型和运行环境镜像	容器镜像服务企业版指引文档 容器镜像服务个人版指引文档
	日志服务-CLS	投递日志用于日志长期归档，日志分析，监控告警	日志服务指引文档
	腾讯云可观测平台-TCOP	在线服务资源使用监控告警	腾讯云可观测平台指引文档
资源组管理	弹性公网 IP	资源组节点机器访问公网资源需要，二者开通一个即可	弹性公网IP指引文档
	NAT 网关		公网 NAT 网关指引文档

	云联网	TI-ONE 平台通过云联网管理资源组节点	云联网指引文档
	腾讯云可观测平台-TCOP	资源组资源使用监控告警	腾讯云可观测平台指引文档

腾讯混元大模型系列

Hunyuan-Large x TI 上手指南

最近更新时间：2025-02-25 21:26:32

总览

腾讯混元宣布最新开源的 MoE 模型 Hunyuan-Large 正式上线，腾讯云 TI 平台第一时间开放接入，为用户和开发者提供一站式精调和部署服务。Hunyuan-Large 拥有 3890 亿总参数量、520 亿激活参数量，并支持 256K 上下文长度，是目前业界参数规模最大、性能领先的开源 MoE 模型。基于 MoE（Mixture of Experts）结构的优越性，混元 Large 在推理速度和参数规模之间取得平衡，显著提升了模型的处理能力。测试结果显示，Hunyuan-Large 在 CMMLU、MMLU、CEval、AGIEval 等多学科评测集以及中英文 NLP 任务、代码、数学等9大维度表现出色，超越 Llama3 和 Mixtral 等主流开源模型。

对应您使用 Hunyuan-Large 的不同场景，TI 平台及腾讯云均能满足您的需求。

场景1：在您的大模型应用接入 Hunyuan-Large 基础模型 API

如果未经精调的 Hunyuan-Large 的效果已能满足您的需求，我们推荐您使用腾讯云提供的 [Hunyuan-Large 公有 API](#)。您可以将 Hunyuan-Large 公有 API 接入您的大模型应用，感受 Hunyuan-Large 基础模型在您的实际应用中产生的作用。
请注意：精调后的专属 Hunyuan-Large 大模型 API 需通过 TI 平台发布。

场景2：基于自有数据精调 Hunyuan-Large 并发布为 API

您可以参考文档 [精调 Hunyuan-Large 实践教程](#) 逐步完成 Hunyuan-Large 模型 SFT 精调 + 部署全流程，包括数据准备、模型精调训练、精调后模型直接发布为 API 等步骤。
基于 TI 平台精调出的专属 Hunyuan-Large 大模型，可以在平台内快速发布为 API，以供您接入应用增强生产力。

您可通过腾讯云提供的镜像下载 Hunyuan-Large 大模型：

模型	下载链接
Hunyuan-A52B-Instruct (Hunyuan-Large-Instruct)	单击下载
Hunyuan-A52B-Pretrain	单击下载
Hunyuan-A52B-Instruct-FP8	单击下载

精调 Hunyuan-Large 实践教程

最近更新时间：2025-06-12 14:27:11

总览

本文将介绍如何使用 [腾讯云 TI 平台](#) 来一步一步完成 Hunyuan-Large 模型的有监督精调（SFT）全流程，本案例使用 [Hunyuan-Large 开源仓库](#) 中附带的约 5000 条汽车领域的 [实体抽取数据](#)（下文称为 car_train 数据集）对 Hunyuan-Large-Instruct 模型做实体抽取任务的有监督精调，以让模型的输出格式更加符合这一特定任务的要求。

训练集数据示例如下（[点击下载完整数据集](#)）：

```
{
  "messages": [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": "请提取下列文章中包含的车系：\nbilstein evo s,黑黄涂装,比b14更舒适的绞牙避震 | 酷乐汽车\n"
    },
    {
      "role": "assistant",
      "content": "无"
    }
  ]
},
{
  "messages": [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": "请提取下列文章中包含的车系：\nbeijingx7燃油版二手车车架号维保查询车辆出险记录是随车不随车主的,这就说明,即使换了投保的保险公司,或是更换车牌号,甚至是出手卖掉,办理车辆过户手续等等,都不能更改和删除车辆的出险记录.\n"
    },
    {
      "role": "assistant",
      "content": "beijingx7"
    }
  ]
},
{
  "messages": [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": "请提取下列文章中包含的车系：\n奥迪q5l作为豪华品牌中最畅销的suv车型也迎来了降价的命运,目前最高优惠8.47万,性价比还是很高的.\n"
    },
    {
      "role": "assistant",
      "content": "奥迪q5l"
    }
  ]
},
{
  "messages": [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": "请提取下列文章中包含的车系：\n“居家暖男” 本田cr-v的变与不变\n"
    },
    {
      "role": "assistant",
      "content": "本田cr-v"
    }
  ]
},
{
  "messages": [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": "请提取下列文章中包含的车系：\nbmw官方认证二手车|骨子里的运动,一如既往\n"
    },
    {
      "role": "assistant",
      "content": "无"
    }
  ]
}
```

下面是一组精调前后模型效果的对比示例（您可以到腾讯云 TI 平台的 Hunyuan-Large [模型体验](#) 页面体验精调前模型效果）：

- 输入："请提取下列文章中包含的车系：\n此外,再配上一幅多辐式的轮毂设计,为该车带来了不错的运动感.\n"
- 精调前输出："根据提供的文章内容，无法直接提取出具体的车系信息。文章中仅描述了车辆的一些设计特征，如“一幅多辐式的轮毂设计”，但并未明确提及车系名称。\n\n因此，无法从该段文字中提取出车系。"
- 精调后输出："无"

前置准备

为了完成本实践，需要先准备如下资源：

数据集

平台已经在 COS 存储中预置了上述 car_train 数据集，您可以在后续步骤中直接使用。

若您需要精调自有数据集，可以在后续的数据准备流程中自行上传，但是具体数据清洗的配置、训练超参等需要参考本最佳实践文档按实际情况调整，整体操作耗时也会有变化。

算力资源

Hunyuan-Large 模型精调需要使用 GPU 算力资源，推荐机型为 HCCPNV6 机型。本文以对 Hunyuan-Large-Instruct 模型做 LoRA 微调为例，需要最少 1 台 HCCPNV6 机型算力资源。若您需要对 Hunyuan-Large-Instruct 模型进行全参数 SFT 精调，推荐准备最少 8 台 HCCPNV6 机型算力资源。

- 自行购买算力：**Hunyuan-Large 模型精调当前支持包年包月算力模式。请参考 [资源组管理](#) 指引，请联系腾讯云客户经理购买 HCCPNV6 机型，然后将已购买的 CVM 添加到 TI 平台资源组中。添加完毕后，即可在新建任务式建模时选择该资源组。
- 申请免费算力：**如您有 Hunyuan-Large 模型精调的免费测试需求，可通过对接的腾讯云客户经理或售前架构师申请免费测试算力；如当前没有腾讯云人员对接，可填写 [问卷](#) 登记。我们会尽快联系您了解业务需求，发放免费测试算力，并指导您进行使用。

可通过下图所示的文档页面右侧浮动栏联系腾讯云客户经理：

文档中心 入门中心 API 中心 SDK 中心 我的反馈

视频服务与安全产品捉虫活动邀您参加! > HOT 搜索相关文档

三 TI-ONE 训练平台

产品动态

产品简介

安全合规

购买指南

快速入门

平台使用概述

平台使用准备

腾讯混元大模型系列

Hunyuan-Large x TI 上手指南

精调 Hunyuan-Large 实践教程

LLM 大模型系列

快速部署和体验内置开源 LLM 大模型

导入和部署自定义 LLM 大模型 (平台内置推理镜像)

精调内置开源大模型

传统 AI 模型系列

Tikit 快速入门

实践教程

LLM 部署及推理

LLM 训练及评测

LLM 和 MLLM 数据标注

文档中心 > TI-ONE 训练平台 > 快速入门 > 腾讯混元大模型系列 > 精调 Hunyuan-Large 实践教程

精调 Hunyuan-Large 实践教程

最近更新时间：2024-11-14 16:08:02

AI摘要

总览

本文将介绍如何使用 腾讯云 TI 平台 来一步一步完成 Hunyuan-Large 模型的有监督精调 (SFT) 全流程。本案例使用 Hunyuan-Large 开源仓库 中附带的约 5000 条汽车领域的 实体抽取数据 (下文称为 car_train 数据集) 对 Hunyuan-Large-Instruct 模型做实体抽取任务的有监督精调, 以让模型的输出格式更加符合这一特定任务的要求。

训练集数据示例如下 (点击下载完整数据集):

```
{
  "messages": [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": "..."
    }
  ]
}
```

下面是一组精调前后模型效果的对比示例 (您可以到腾讯云 TI 平台的 Hunyuan-Large 模型体验 页面体验精调前模型效果):

- 输入: "请提取下列文章中包含的车系: \n此外,再配上一幅多辐式的轮毂设计,为该车带来了不错的运动感。"
- 精调前输出: "根据提供的文章内容,无法直接提取出具体的车系信息。文章中仅描述了车辆的一些设计特征,如"一幅多辐式的轮毂设计",但并未明确提及车系名称。因此,无法从该段文字中提取出车系。"
- 精调后输出: "无"

前置准备

为了完成本实践,需要先准备如下资源:

数据集

本文目录:

- 总览
- 前置准备
 - 数据集
 - 算力资源
 - 文件存储
- 实践教程
 - 步骤一: 数据准备
 - 步骤二: 模型精调训练
 - 步骤三: 精调后模型部署
- 附录

售前咨询

专属客户经理在线解答,提供专业的上云方案

售后支持

7*24获取专业工程师的帮助,快速解决您的问题

查看更多联系、反馈方式 >

文档反馈

文档活动

文件存储

模型精调过程中,数据集和模型 checkpoint 存储都依赖 CFS,请到 [CFS 控制台](#) 开通 CFS。备注说明:本文以 CFS 为例进行详细说明,精调内置大模型时也支持挂载 GooseFSx (当机器来源为从 CVM 机器中选择) 作为用户自有数据;同时创建开发机也支持挂载 GooseFSx。

由于 Hunyuan-Large 模型文件非常大,为保证训练保存 checkpoint 及后续使用模型部署推理服务和启动评测的速度,推荐使用 **Turbo 标准型存储**,CFS 实例区别详见 [存储类型及规格](#)。请保证购买的 CFS 实例与上述算力资源机器的网络互通。

实践教程

下面我们开始具体的实践教程,本实践主要包括数据准备、模型精调训练、精调后模型部署几个步骤:

- 数据准备:** 对原始的数据集进行格式转换,将处理好的数据存放在 CFS 或者 GooseFSx 中,便于后续启动精调训练。使用示例数据按实践教程操作本步骤耗时约5分钟;
- 模型精调训练:** 为了便于算力资源受限的用户也能在腾讯云 TI 平台体验精调 Hunyuan-Large 模型,本示例对 Hunyuan-Large-Instruct 模型做 LoRA 微调。使用示例数据按实践教程操作本步骤耗时约4~5小时;
- 精调后模型部署:** 将精调后的模型部署为在线推理服务,对外提供 API 调用能力,您可以进一步评估模型效果或接入自有的应用。按实践教程操作,本步骤耗时约10~60分钟 (首次部署耗时会偏长,后续视实际CFS存储的性能而定);

步骤一: 数据准备

数据准备工作会用到腾讯云 TI 平台的【数据中心-数据构建】及【训练工坊-开发机】功能,具体步骤如下:

1.1 创建数据构建开发机实例

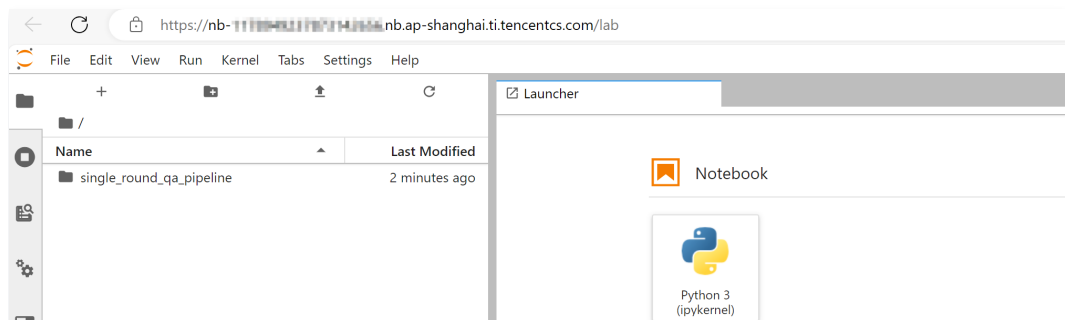
进入[数据中心](#) > [数据构建](#) 页面,单击**新建任务**,输入任意任务名称,数据构建 pipeline 选择【有监督-单轮问答-pipeline】。



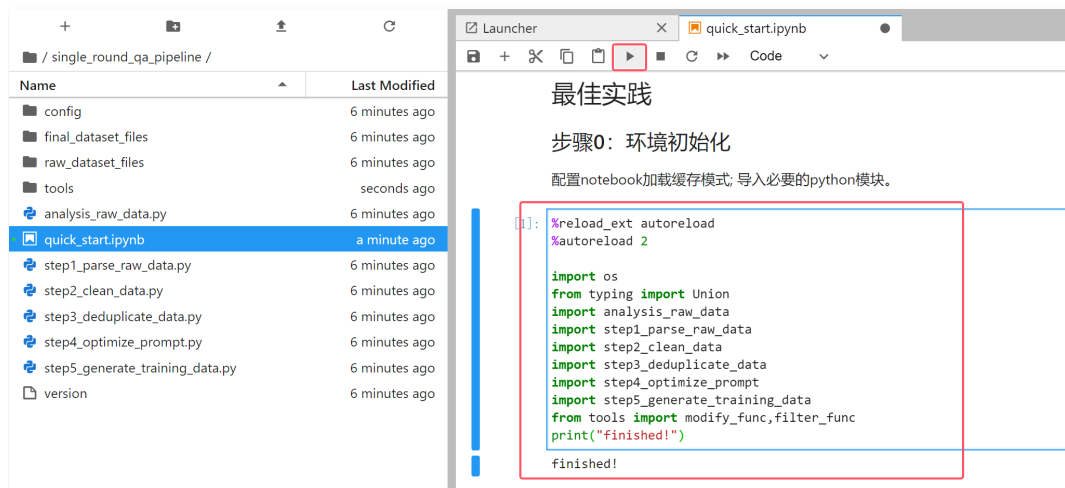
单击跳转到对应开发机按钮。首次跳转开发机会进入新建开发机页面，这里示例使用按量计费 4C8G 的实例规格，若您有资源组机器，也可以使用包年包月模式选择对应机器，数据构建流程不需要使用 GPU，实例的 CPU 核数越多，数据构建的数据去重步骤速度越快，其他影响不是很大。

1.2 在开发机中进行数据处理

等待开发机实例启动，大概需要2分钟左右，启动完成后实例状态会变为“运行中”，此时在右侧操作栏单击**打开**，会跳转到实例的 Jupyterlab 页面。

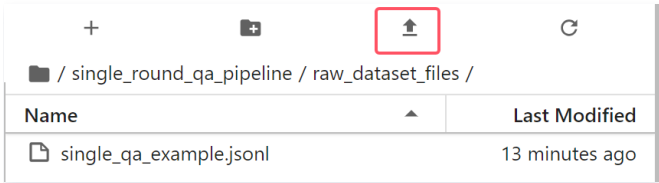


双击进入 `single_round_qa_pipeline` 目录，打开 `quick_start.ipynb` 文件，可以打开数据构建指引。按照指引一步一步操作，首先单击“步骤0：环境初始化”对应的代码块，单击运行：



1.2.1 上传数据集

接下来需要上传原始数据集，若您的数据集在本地，且文件不是非常大，可以双击 `raw_dataset_files` 目录，单击上传按钮将本地的数据集上传上去：



这里我们以 Hunyuan-Large 开源仓库里汽车领域的实体抽取数据为例，把 `car_train.jsonl` 上传到开发机的 `raw_dataset_files` 目录里。回到 `quick_start.ipynb`，在“步骤1：原始数据解析”这一步中，修改 `input_path` 为刚刚的数据集文件名，运行后可以看到原始数据集的样本数及 json 结构信息及5条采样数据示例，如图：

步骤1.1 原始数据采样

该步骤内置脚本的功能是：读取 `step1_in_file` 的原始数据，自动统计和采样指定数目的样例数据，输出对应数据的schema，并展示在下方，这样用户可以清晰了解输入数据的格式。TI 平台的单轮对话数据构建 pipeline 要求的每条样本输入数据格式定义如下，您可跟进此处展示的schema确定数据转换的方式。

```
{ "system": "xxxx", "question": "xxxx", "answer": "xxxx" }
```

[31]:

```
# 采样数据
# 原始数据解析输入的原始文件路径，建议原始输入统一用 "raw_dataset_files" 文件夹进行管理；
# 输入支持文件和文件夹
input_path = "car_train.jsonl" # 单个文件

step1_in_file = os.path.join("raw_dataset_files", input_path)
dataset_name = os.path.splitext(os.path.basename(input_path))[0].replace(".", "_")
print("数据集名称:{}".format(dataset_name))

# 使用同一目录下的 analysis_raw_data.py 中定义的函数进行原始数据采样
analysis_raw_data.sample_raw_data_single_round_qa(step1_in_file, sample_num = 5)
```

数据集名称:car_train
开始处理数据文件raw_dataset_files/car_train.jsonl ...
QA-数据采样
数据集名称:raw_dataset_files/car_train.jsonl, 数据总量:4608, 采样数据量:5
数据json schema格式如下:



可以看到数据集均为带 `system` 的单轮对话数据 `messages[0]["content"]` 字段对应平台要求的 `system` 字段，`messages[1]["content"]` 字段对应平台要求的 `question` 字段，`messages[2]["content"]` 字段对应平台要求的 `answer` 字段。我们相应的修改对应代码：

```
question = item["messages"][1]["content"]
answer = item["messages"][2]["content"]
system = item["messages"][0]["content"]
```

如图所示：

```
281: # 解析数据

# 解析后的数据输出路径, jsonl 格式, 建议输出统一用 "intermediate_dataset_files" 文件夹进行管理
step1_out_file = os.path.join(
    "intermediate_dataset_files", "01_{}_parsed.jsonl".format(dataset_name)
)

# 您可以重写这个函数来实现对不同json格式的解析, 对于您的输入文件, 我们会解析成多行的json, 并把json格式转成python dict格式。
# 在该函数中, 您根据采样的json schema把对应格式的dict转成tione平台指定的格式{"system": "xxxx", "question": "xxxx", "answer": "xxxx"}即可。
def convert_dict_to_tione_format(item: dict) -> dict:
    try:
        question = item["messages"][1]["content"]
        answer = item["messages"][2]["content"]
        # 样本构造: 获取系统角色字段"system", 具体的内容用户可按需修改
        system = item["messages"][0]["content"]
        # 构造tione要求的样本格式
        convert_dict = {
            "system": system,
            "question": question,
            "answer": answer,
        }
    except:
        # 数据解析异常
        return None
    return convert_dict
```

运行后会输出转换后的数据集文件及一些数据分布信息:

```
开始处理数据文件raw_dataset_files/car_train.jsonl ...
100% [██████████] 4605/4605 [00:00<00:00, 146094.91it/s]
总样本数量: 4605
解析异常样本数量: 0
解析完成输出文件:/home/tione/notebook/nb-1178949237873142656/single_round_qa_pipeline/intermediate_dataset_files/01_car_train_parsed.jsonl
开始执行统计...
100% [██████████] 4605/4605 [00:00<00:00, 51482.30it/s]
有监督训练数据prompt部分平均字符长度:63.85559174809989
有监督训练数据response部分平均字符长度:3.7040173724212813
```

1.2.2 训练格式生成

由于这里的示例数据集已经预先经过数据清洗和去重等工作, 因此我们跳过数据构建中的步骤2、3、4, 直接来到最后一步训练格式生成, 这里我们修改这一步的输入数据为上面第一步的输出(如下图红框所示), 执行完后可以拿到最终处理好的数据集在 CFS 或者 GooseFSx 上的路径, 我们复制一下这个路径, 用于后续模型精调训练时输入用户自有数据的源路径。

步骤5: 训练格式生成

该步骤内置脚本的功能是: 将前序步骤已经处理好的有监督-单轮问答的数据转换成数据构建 pipeline 的输出数据格式(即后续的训练数据格式), 并输出必要的训练数据统计分析结果(数据集数据量、prompt部分平均长度、response部分平均长度)。

原始代码逻辑详见 `step5_generate_training_data.py` (一般无特殊情况下, 用户无需修改 python 文件的原始代码)。

数据构建 pipeline 输出数据每条样本的格式如下(即训练数据格式):

```
{
    "system": "xxxx",
    "conversation": [{"prompt": "xxxx", "response": "xxxx"}]
}
```

```
# 训练数据生成输入 jsonl 文件路径, 一般是上一步的输出
step5_in_file = step1_out_file

# 训练数据生成输出 jsonl 文件路径, 建议最终的输出统一用"final_dataset_files"文件夹进行管理
step5_out_file = os.path.join(
    "final_dataset_files", "{}_for_training.jsonl".format(dataset_name)
)

# 使用同一目录下的 step5_generate_training_data.py 中定义的函数进行训练数据生成
# 其中参数tokenizer_model_path可以指定计算token的模型路径, 默认为python/Len, 表示不使用大模型, 直接计算字符串的长度;
step5_generate_training_data.generate_training_data_single(step5_in_file, step5_out_file, tokenizer_model_path = tokenizer_model_path)

100% [██████████] 4605/4605 [00:00<00:00, 137227.49it/s]
处理完成: car_train_for_training.jsonl
数据集数据量: 4605
有监督训练数据prompt部分平均字符长度:63.85559174809989
有监督训练数据response部分平均字符长度:3.7040173724212813
可直接复制(按Ctrl+C键复制)该路径去启动训练任务(映射到挂载cfs路径):/nb-1178949237873142656/single_round_qa_pipeline/final_dataset_files
```

步骤二：模型精调训练

模型精调训练步骤会用到腾讯云 TI 平台的【大模型广场】和【训练工坊-任务式建模】功能, 具体步骤如下:

2.1 启动训练任务

进入 **大模型广场** 菜单, 单击混元大模型下的 Hunyuan-Large 卡片。



单击**新建训练任务**，会自动跳转到【训练工坊-任务式建模】的新建任务页面，按需填写任务名称。如果使用包年包月的资源组机器，单节点的GPU、CPU、内存资源尽量用满（H20单节点整机可用CPU380核，内存2214GB，若整机为空闲状态，推荐填写CPU380核，内存2214GB；若整机资源被部分占用，可以点击资源组下方可用卡数的**查看详情**链接，查看资源组内每个节点的可用资源），如下图所示：

基本信息

任务名称 *

Hunyuan-Large精调实践-汽车实体抽取-Lora-单机

请输入不超过60个字符，仅支持中英文、数字、下划线"_"、短横"-", 只能以中英文、数字开头

地域 *

上海

训练镜像 *

内置大模型 / Hunyuan-Large / Hunyuan-Large-Instruct

训练模式 *

DDP

计费模式 *

按量计费

包年包月

资源组 *

TIONE-H20

TI-ONE 仅收取软件费用，算力将使用您已购买的机器。[产品定价](#)
HCC-H20可用卡数：总计4/整机0/非整机4 [查看详情](#)

资源申请 *

卡型号

HCC-H20

单节点GPU

-

8

+

卡

该机型只支持填写GPU整数卡

单节点CPU *

-

380.0

+

核

单节点内存 *

-

2214.00

+

G

节点数 *

-

1

+

个

存储路径设置：

1. 训练数据：这里不使用平台内置数据，单击平台内置数据这一条最右侧的删除，然后添加存储类型为 CFS Turbo 或者 GooseFSx 、用途为"用户自有数据"的一行，CFS 或者 GooseFSx 选择前面数据构建用的 CFS Turbo 或者 GooseFSx，源路径填写 [训练格式生成](#) 章节最后复制的路径；
2. 训练输出：这里我们使用 CFS Turbo 存储模型输出，因此先单击原有的训练输出最右侧的删除，然后添加存储类型为 CFS Turbo 、用途为"训练输出"的一行，CFS 选择准备好的 CFS Turbo 实例，源路径填写一个全新的目录或一个已存在的空目录；

任务配置

存储路径设置 ①

若您需挂载多个存储路径，需要确保选择的存储实例（CFS、CFS Turbo或者EMR(HDFS)）在同一个VPC或者同一个云联网下，系统当前会默认打通选择的第一个存储实例所在的VPC网络或云联网网络

存储类型	用途	CFS文件系统	源路径 ①	容器挂载路径 ①	操作
CFS	平台内置代码	平台CFS	/code	/opt/ml/code	删除
CFS	平台内置数据	平台CFS	/data	/opt/ml/input/data/train	删除
CFS	平台内置模型	平台CFS	Hunyuan-Large/Hunyuan-L...	/opt/ml/pretrain_model	删除
CFS	训练输出	请选择	/	/opt/ml/output/data	删除

任务配置

存储路径设置 ①

请确保您选择的存储实例（CFS、CFS Turbo、EMR(HDFS)或者GooseFSx）和纳管资源组的节点网络互通，其中GooseFSx仅支持挂载一个实例

存储类型	用途	CFS文件系统	源路径 ①	容器挂载路径 ①	操作
CFS	平台内置代码	平台CFS	/code	/opt/ml/code	删除
CFS	平台内置模型	平台CFS	Hunyuan-Large/Hunyuan-L...	/opt/ml/pretrain_model	删除

存储类型	用途	CFS Turbo文件系统	源路径 ①	容器挂载路径 ①	操作
CFS Turbo	训练输出	ti-one-hunyuan (cfs- ti-one-hunyuan)	/output/hunyuan-large-lora-test1	/opt/ml/output/data	删除
CFS Turbo	用户自有数据	ti-one-hunyuan (cfs- ti-one-hunyuan)	/nb-1178949237873142656/single	/opt/ml/input/data/train	删除

[查看训练数据格式](#)

调优参数配置建议：

由于这里我们以单机8卡作为示例，全局 BatchSize 较小，这里我们稍稍调大 GradientAccumulationSteps 为 2，启动训练。

代码包 ①

选择文件 清空

请选择对象存储 COS 中的文件

启动命令 ① *

```
1 cd /opt/ml/code; bash start.sh
```

30/8192

训练输出 ①

选择目录 清空

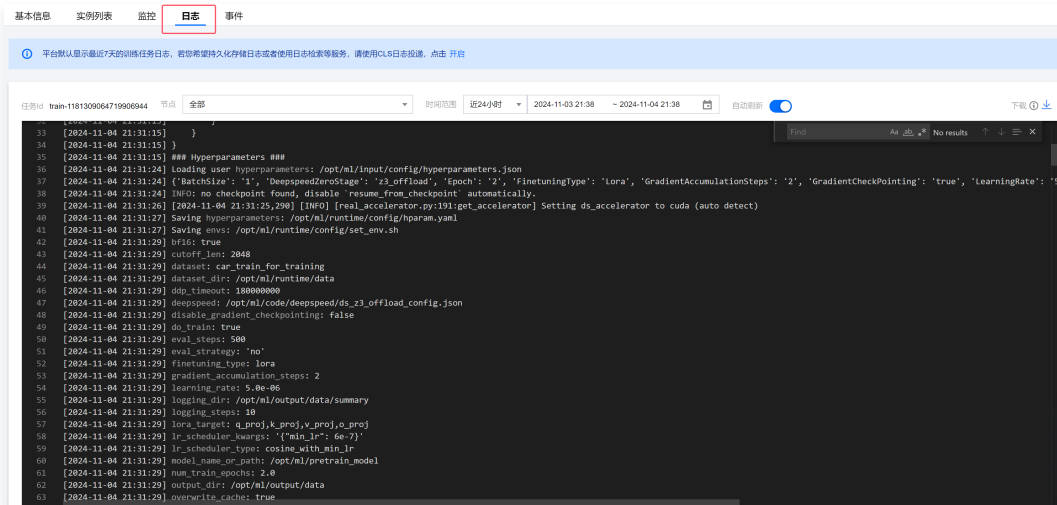
调优参数 ①

```
1 {
2   "Epoch": "2",
3   "BatchSize": "1",
4   "LearningRate": "5e-6",
5   "Step": "500",
6   "FinetuningType": "Lora",
7   "MaxSequenceLength": "2048",
8   "GradientAccumulationSteps": "2",
9   "GradientCheckpointing": "true",
10  "DeepspeedZeroStage": "z3_offload",
11  "ResumeFromCheckpoint": "true"
```

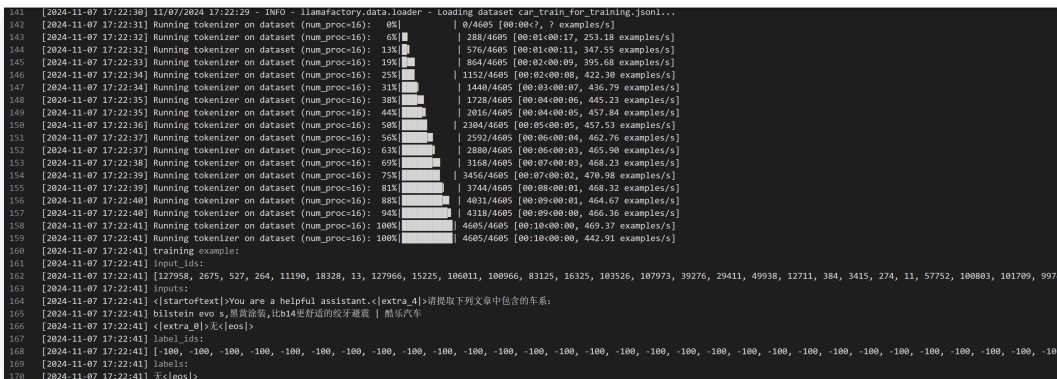
2.2 训练状态观察

训练任务启动后，建议先单击日志观察训练状态（若您启动的是多机分布式训练任务，建议日志页面"节点"下拉框选择 master-0 节点，以免日志过多），训练任务启动后会有以下几个阶段：

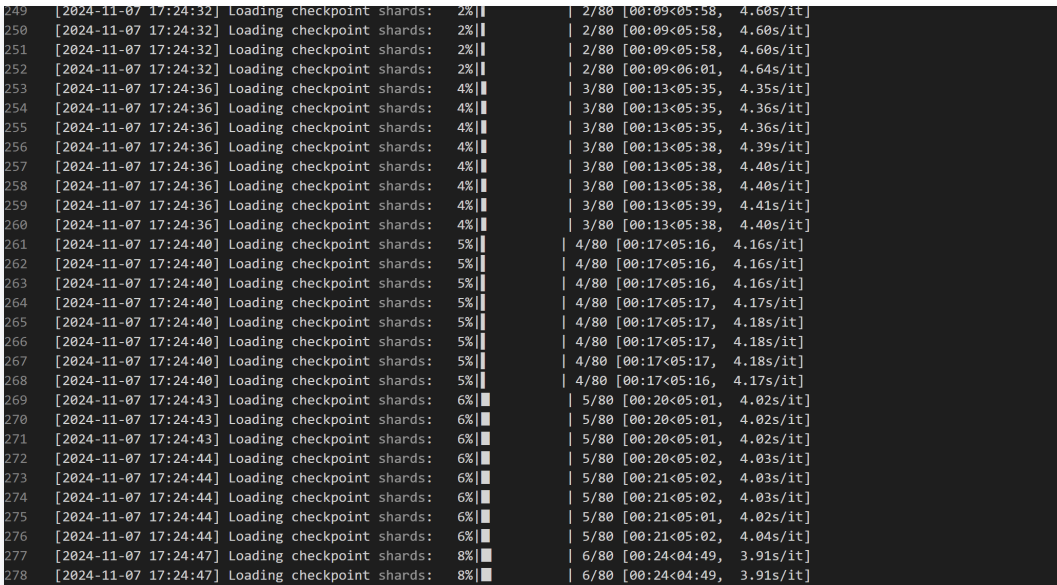
1. 加载用户指定的调优参数，打印最终任务使用的超参列表，这里的超参兼容开源 **LLaMA-Factory** 框架，若您非常清楚实际超参的含义及修改后果，也可以在平台的调优参数中添加平台未预置但是 LLaMa-Factory 0.8.3 版本中有的超参来覆盖最终训练使用的超参；



2. 加载数据集，并使用模型的Tokenizer和对话模板对数据进行预处理，打印部分数据预处理后的样本；



3. 加载预训练模型权重，这一步耗时可能偏长，我们为Hunyuan-Large体验机型提供了缓存优化能力；



4. 准备训练，打印 DeepSpeed 一些参数信息及训练任务信息，包括要训练的样本数量、全局 batch size、总迭代步数等信息；

```
1190 [2024-11-07 17:29:14] "stage": 3,
1191 [2024-11-07 17:29:14] "offload_optimizer": {
1192 [2024-11-07 17:29:14]   "device": "cpu",
1193 [2024-11-07 17:29:14]   "pin_memory": true
1194 [2024-11-07 17:29:14] },
1195 [2024-11-07 17:29:14] "offload_param": {
1196 [2024-11-07 17:29:14]   "device": "cpu",
1197 [2024-11-07 17:29:14]   "pin_memory": true
1198 [2024-11-07 17:29:14] },
1199 [2024-11-07 17:29:14] "overlap_comm": true,
1200 [2024-11-07 17:29:14] "contiguous_gradients": true,
1201 [2024-11-07 17:29:14] "sub_group_size": 1.000000e+09,
1202 [2024-11-07 17:29:14] "reduce_bucket_size": 4.096000e+07,
1203 [2024-11-07 17:29:14] "stage3_prefetch_bucket_size": 3.686400e+07,
1204 [2024-11-07 17:29:14] "stage3_param_persistence_threshold": 6.400000e+04,
1205 [2024-11-07 17:29:14] "stage3_max_live_parameters": 1.000000e+09,
1206 [2024-11-07 17:29:14] "stage3_max_reuse_distance": 1.000000e+09,
1207 [2024-11-07 17:29:14] "stage3_gather_16bit_weights_on_model_save": true
1208 [2024-11-07 17:29:14] },
1209 [2024-11-07 17:29:14] "steps_per_print": inf
1210 [2024-11-07 17:29:14] }
1211 [2024-11-07 17:29:14] [INFO|trainer.py:2134] 2024-11-07 17:29:13.992 > ***** Running training *****
1212 [2024-11-07 17:29:14] [INFO|trainer.py:2135] 2024-11-07 17:29:13.992 >   Num examples = 4,605
1213 [2024-11-07 17:29:14] [INFO|trainer.py:2136] 2024-11-07 17:29:13.992 >   Num Epochs = 2
1214 [2024-11-07 17:29:14] [INFO|trainer.py:2137] 2024-11-07 17:29:13.992 >   Instantaneous batch size per device = 1
1215 [2024-11-07 17:29:14] [INFO|trainer.py:2140] 2024-11-07 17:29:13.992 >   Total train batch size (w. parallel, distributed & accumulation) = 16
1216 [2024-11-07 17:29:14] [INFO|trainer.py:2141] 2024-11-07 17:29:13.992 >   Gradient Accumulation steps = 2
1217 [2024-11-07 17:29:14] [INFO|trainer.py:2142] 2024-11-07 17:29:13.992 >   Total optimization steps = 576
1218 [2024-11-07 17:29:14] [INFO|trainer.py:2143] 2024-11-07 17:29:14.007 >   Number of trainable parameters = 16,711,680
1219 [2024-11-07 17:29:15] exit | 0/576 [00:00<, 7it/s]
```

5. 开始训练，打印训练迭代步数、loss、epoch 等进度信息，这里默认每 10 步打印一次训练信息日志，示例中大概每一步迭代耗时在 25 秒左右；

```
[2024-11-07 17:42:12] 5%| | 30/576 [12:57<3:48:42, 25.13s/it]
[2024-11-07 17:42:12] {'loss': 4.0639, 'grad_norm': 3.7666070061293953, 'learning_rate': 4.980780952519346e-06, 'epoch': 0.1}
[2024-11-07 17:42:37] 5%| | 31/576 [13:22<3:48:13, 25.12s/it]
[2024-11-07 17:43:02] 6%| | 32/576 [13:47<3:47:44, 25.12s/it]
[2024-11-07 17:43:27] 6%| | 33/576 [14:12<3:47:13, 25.11s/it]
[2024-11-07 17:43:53] 6%| | 34/576 [14:38<3:46:49, 25.11s/it]
[2024-11-07 17:44:18] 6%| | 35/576 [15:03<3:46:34, 25.13s/it]
[2024-11-07 17:44:43] 6%| | 36/576 [15:28<3:46:14, 25.14s/it]
[2024-11-07 17:45:08] 6%| | 37/576 [15:53<3:45:39, 25.12s/it]
[2024-11-07 17:45:33] 7%| | 38/576 [16:18<3:45:07, 25.11s/it]
[2024-11-07 17:45:58] 7%| | 39/576 [16:43<3:44:58, 25.14s/it]
[2024-11-07 17:46:23] 7%| | 40/576 [17:08<3:44:37, 25.14s/it]
[2024-11-07 17:46:23] {'loss': 3.6285, 'grad_norm': 1.8655208667210756, 'learning_rate': 4.961485054959454e-06, 'epoch': 0.14}
[2024-11-07 17:46:49] 7%| | 41/576 [17:34<3:45:16, 25.26s/it]
[2024-11-07 17:47:14] 7%| | 42/576 [17:59<3:44:32, 25.23s/it]
[2024-11-07 17:47:39] 7%| | 43/576 [18:24<3:43:50, 25.20s/it]
[2024-11-07 17:48:04] 8%| | 44/576 [18:49<3:43:20, 25.19s/it]
[2024-11-07 17:48:30] 8%| | 45/576 [19:15<3:42:55, 25.19s/it]
[2024-11-07 17:48:55] 8%| | 46/576 [19:40<3:42:25, 25.18s/it]
[2024-11-07 17:49:20] 8%| | 47/576 [20:05<3:41:59, 25.18s/it]
[2024-11-07 17:49:45] 8%| | 48/576 [20:30<3:41:32, 25.17s/it]
[2024-11-07 17:50:10] 9%| | 49/576 [20:55<3:41:01, 25.16s/it]
[2024-11-07 17:50:35] 9%| | 50/576 [21:20<3:40:40, 25.17s/it]
[2024-11-07 17:50:35] {'loss': 3.3835, 'grad_norm': 2.0027285083783783, 'learning_rate': 4.935624798328384e-06, 'epoch': 0.17}
[2024-11-07 17:51:01] 9%| | 51/576 [21:46<3:40:22, 25.19s/it]
```

训练到达指定一定步数或训练完成时，会保存 checkpoint，保存的 checkpoint 可以用于后续启动推理和评测等，下图为保存 checkpoint 的日志。

```
854 [2024-11-05 02:57:58] 88%| | 1,300/576 [13:40:45<3:11:2, 26.70s/it]
850 [2024-11-05 02:54:38] {'loss': 0.3287, 'grad_norm': 0.5023872311742362, 'learning_rate': 7.651281289948403e-07, 'epoch': 1.75}
851 [2024-11-05 02:54:38] [INFO|trainer.py:3581] 2024-11-05 02:57:46.917 > Saving model checkpoint to /opt/ml/output/data/checkpoint-500
852 [2024-11-05 02:57:47] [INFO|trainer.py:3581] 2024-11-05 02:57:46.917 > Saving model checkpoint to /opt/ml/output/data/checkpoint-500
853 [2024-11-05 02:57:47] /usr/local/lib/python3.10/dist-packages/peft/utils/save_and_load.py:195: UserWarning: Could not find a config file in /opt/ml/pretrain_model - will assume that the vocabulary was
854 [2024-11-05 02:57:47] warnings.warn(
855 [2024-11-05 02:57:48] [INFO|tokenization_utils_base.py:2702] 2024-11-05 02:57:47.140 > tokenizer config file saved in /opt/ml/output/data/checkpoint-500/tokenizer_config.json
856 [2024-11-05 02:57:48] [INFO|tokenization_utils_base.py:2711] 2024-11-05 02:57:47.167 > Special tokens file saved in /opt/ml/output/data/checkpoint-500/special_tokens_map.json
857 [2024-11-05 02:57:56] [2024-11-05 02:57:55.593] [INFO] [logging.py:96:log_dist] [Rank 0] [Torch] Checkpoint global_step500 is about to be saved!
858 [2024-11-05 02:57:56] /usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py:1877: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/ma
859 [2024-11-05 02:57:56] warnings.warn(
860 [2024-11-05 02:57:56] /usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py:1877: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/ma
861 [2024-11-05 02:57:56] warnings.warn(
862 [2024-11-05 02:57:56] /usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py:1877: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/ma
863 [2024-11-05 02:57:56] warnings.warn(
864 [2024-11-05 02:57:56] /usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py:1877: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/ma
865 [2024-11-05 02:57:56] warnings.warn(
866 [2024-11-05 02:57:56] /usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py:1877: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/ma
867 [2024-11-05 02:57:56] warnings.warn(
868 [2024-11-05 02:57:56] /usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py:1877: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/ma
869 [2024-11-05 02:57:56] warnings.warn(
870 [2024-11-05 02:57:56] /usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py:1877: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/ma
871 [2024-11-05 02:57:56] warnings.warn(
872 [2024-11-05 02:57:56] /usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py:1877: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/ma
873 [2024-11-05 02:57:56] warnings.warn(
874 [2024-11-05 02:57:56] [2024-11-05 02:57:55.592] [INFO] [logging.py:96:log_dist] [Rank 0] Saving model checkpoint: /opt/ml/output/data/checkpoint-500/global_step500/zero_pp_rank_0_mp_rank_00_model_state
875 [2024-11-05 02:57:56] [2024-11-05 02:57:55.592] [INFO] [torch_checkpoint_engine.py:21:save] [Torch] Saving /opt/ml/output/data/checkpoint-500/global_step500/zero_pp_rank_0_mp_rank_00_model_states.pt...
876 [2024-11-05 02:57:56] [2024-11-05 02:57:55.621] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /opt/ml/output/data/checkpoint-500/global_step500/zero_pp_rank_0_mp_rank_00_model_states.pt...
877 [2024-11-05 02:57:56] [2024-11-05 02:57:55.660] [INFO] [torch_checkpoint_engine.py:21:save] [Torch] Saving /opt/ml/output/data/checkpoint-500/global_step500/rf16_zero_pp_rank_0_mp_rank_00_optim_states.pt...
878 [2024-11-05 02:57:56] [2024-11-05 02:57:55.824] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /opt/ml/output/data/checkpoint-500/global_step500/rf16_zero_pp_rank_0_mp_rank_00_optim_states.pt...
879 [2024-11-05 02:57:56] [2024-11-05 02:57:55.854] [INFO] [engine.py:3488:save_zero_checkpoint] zero checkpoint saved /opt/ml/output/data/checkpoint-500/global_step500/rf16_zero_pp_rank_0_mp_rank_00_optim_states.pt...
880 [2024-11-05 02:57:56] [2024-11-05 02:57:55.882] [INFO] [torch_checkpoint_engine.py:31:commit] [Torch] Checkpoint global_step500 is ready now!
```

若您希望可视化地查看训练任务的各项指标，请参考附录 [使用 Tensorboard 可视化训练过程](#)。

步骤三：精调后模型部署

本步骤会用到腾讯云 TI 平台的【模型服务-在线服务】功能，具体步骤如下：

3.1 部署服务

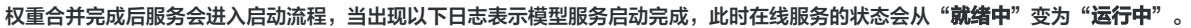
进入模型服务 > [在线服务](#) 菜单，单击新建服务按钮：

服务名称可以自定义，其中模型来源选择 CFS 或者 GooseFSx，并选中模型存储的 CFS 或者 GooseFSx 实例，路径按实际训练输出路径填写，填写到 checkpoint 这一级目录，运行环境选择内置的 LLM / angel-vllm(2.1) 镜像，最低资源需求（按默认开启NF4量化）：内存 1TB，GPU 显存 320 GB，例如 A100 * 8 卡或更大显存的机型。

您可以使用默认配置启动推理服务，若需要了解更多 angel-vllm 推理镜像的特性和高级设置使用方法，可以参考附录[推理镜像特性说明](#)。具体示例如下图所示：

请确保您选择的存储实例（CFS或者GooseFSx）和纳管资源组的节点网络互通

由于本文中使用的精调方式是 LoRA，因此默认保存的 checkpoint 并不是完整模型权重，在推理时需要和原始权重进行合并。通过平台首次启动 LoRA 模型时，服务会进行权重的自动合并，并将合并后的模型保存在模型目录的 `merged_model` 子目录下，后续重新用该 checkpoint 启动推理服务的话，会自动使用合并后的权重，无需重新合并。如下图日志表示正在合并权重：



回到上级菜单，单击服务的【服务调用】标签页面：



然后输入请求体，接口格式兼容目前流行的 OpenAI Chat Completions 接口，如图所示，我们可以拿 Hunyuan-Large 开源项目里带的 `car_test.json` 测试集抽几条样本试验一下，例如请求体为：

版权所有：腾讯云计算（北京）有限责任公司

的运动感。\\n"}]]}

如图所示：

接口信息

接口调用地址

https://sw.gw.ap-shanghai-8.tencentcloudapi.com/v1/chat/completions

服务类型

HTTP

请求方法

POST

调用方式(命令行)

curl -X POST https://sw.gw.ap-shanghai-8.tencentcloudapi.com/v1/chat/completions -H 'Content-Type: application/json' -d '{
 "messages": [{"role": "user", "content": "请提取下列文章中包含的车系：\\n此外，再配上一辆多轴式的轮毂设计，为该车带来了不错的运动感。\\n"}]}'
若服务开启了鉴权，请参考[文档](#) [指引](#)调用

调用方式(在线测试)

请求体(Request Body 600KB 内)

1 [{"messages":[{"role":"user","content":"请提取下列文章中包含的车系：\\n此外，再配上一辆多轴式的轮毂设计，为该车带来了不错的运动感。\\n"}]}]

请求响应(Response)

```
12  "object": "chat.completion",
13  "created": 1738981641,
14  "model": "hunyuan-large-chat-v1",
15  "choices": [
16    {
17      "index": 0,
18      "message": {
19        "role": "assistant",
20        "content": "无",
21        "tool_calls": []
22      },
23      "logprobs": null,
24      "finish_reason": "stop",
25      "stop_reason": 127960
26    }
27  ],
28  "usage": {
29    "prompt_tokens": 38,
30    "total_tokens": 40,
31    "completion_tokens": 2
32  }
```

发送请求

您也可以直接使用 OpenAI SDK 或者其他大模型应用框架如 Dify、LangChain 等接入腾讯云 TI 平台在线服务的 API，endpoint 填写腾讯云 TI 平台的接口调用地址后面加 /v1 即可。若您需要开启类似 OpenAI API key 的鉴权，可以在服务启动时增加环境变量 VLLM_API_KEY 为您需要设置的 API_KEY，并在调用时指定即可。

附录

使用 Tensorboard 可视化训练过程

有可能直接通过训练任务的日志观察训练指标不够直观，此时我们也可以使用平台提供的 Tensorboard 功能来更加可视化地观察训练过程，指引如下：
单击训练任务右侧操作一栏中的 Tensorboard 按钮。

名称	训练镜像	训练模式	计费模式	占用资源 ①	标签	状态	运行时长	创建者	训练开始时间 ②	更新时间 ③	监控	操作
Hunyuan-Large精调实践-汽车... train-1183356174146071552	内置大模型 Hunyuan-Large-Hunyuan-La rge-Instruct	DDP	包年包月 TIONE-H20	算力规格：380C2214G HCC- H20* 节点数量：1 个		已完成	4小时23分13秒	100039273032	2024-11-07 17:21:21	2024-11-07 17:17:31		复制 Tensorboard 启动 更多

查看Tensorboard

选择summary输出存储

CFS

选择存储路径

tion-hunyuan (cfs-) | /output/hunyuan-large-lora-test1/

summary目录 ①

/opt/ml/output/data/summary

访问地址

[点击跳转](#)

单击跳转，稍等片刻会跳转到任务对应的 tensorboard 页面，如下图所示：



可以观察任务的学习率、loss 等变化曲线，若 loss 收敛情况良好，一般没有太大问题。若您的数据在实际精调时 loss 收敛情况不好，可能需要具体分析，可以参照下面的 [训练调参指引](#) 进行一些调参优化。

训练调参指引

Hunyuan-Large 模型学习能力很强，在小规模训练数据上可能会有过拟合问题。平台默认参数为推荐参数，可以先跑 2 个 epoch 观察下训练 loss 曲线收敛情况再做调整，建议如下：

超参调整建议

- Epoch: 训练轮次推荐取值 2~3, MoE 模型拟合能力强, 建议不要超过 5, 避免过拟合;
- BatchSize: 默认为 1, 不建议随意调大, 否则容易出现显存不足问题, 显存富余较多时再考虑调整;
- LearningRate: 学习率建议在 $5e-6 \sim 1e-4$ 范围内调整, 可以视实际情况调整:
 - 如果训练 loss 收敛太快, 或者出现明显震荡, 说明可能学习率偏大;
 - 如果训练 loss 收敛太慢, 或者收敛值很高, 说明可能学习率偏小;
 - 如果训练 loss 收敛值小于 0.1, 可能存在过拟合风险, 建议尝试调小学习率或 Epoch;
 - 如果训练 loss 收敛值大于 1, 可能训练还不够充分, 建议尝试调大学习率或 Epoch;
- Step: 表示每隔多少步迭代保存一次 checkpoint, 建议根据训练的总迭代步数调整;
- FinetuningType: 若机器资源很少, 选择 "Lora"; 若机器资源足够, 希望达到更高的模型精调性能上限或者训练数据集较大, 推荐选择 "Full";
- MaxSequenceLength: 建议根据实际训练集数据 token 长度调整, 超长的数据会被截断影响效果;
- GradientAccumulationSteps: 梯度累积步数会影响全局 BatchSize, 全局 BatchSize = 训练卡数 * GradientAccumulationSteps * 单卡 BatchSize, 建议全局 BatchSize 在 32 或 64 左右, 同时尽量保证总迭代步数不要太少 (推荐 500 步以上);

训练数据建议

- 如果训练起始 loss 非常高, 例如大于 10, 可能是数据存在问题, 或者使用的 Tokenizer 或对话模板错误等, 建议仔细排查; 偏高的起始 loss 一般说明训练数据与基座模型的训练数据区别较大;
- 为避免过拟合, 增加样本多样性, 可以考虑加入一些通用数据一起训练, 如平台数据构建内置的一些算法场景数据;
- 本文使用的数据仅为示例数据, 实际业务中使用的训练数据可能更加复杂, 可以综合结合数据的特点及其他开源模型的精调经验来调整训练超参;

推理镜像特性说明

内置 angel-vllm(2.1) 推理镜像使用了腾讯自研优化的 Tilearn Angel-vLLM 框架进行了推理加速, 其在开源 vLLM 框架基础上额外支持了以下亮点特性:

- **NF4 (NormalFloat) 量化**: 相比开源 vllm 的实现, 无需做模型转换, 在保持量化精度的同时充分利用 GPU 资源实现了推理性能的大幅提升, 且对 Hunyuan-Large 这样的混合专家模型 (MoE) 进行了特定优化;
- **Lookahead并行解码**: 相比开源实现, 无需额外的小模型或模型头, 可以做到并行解码的结果和非并行解码的结果一致, 对于输出文本中有大部分在输入文本中都出现过 (例如 RAG 场景), 或是大量请求中有相似请求或答案的情况下有明显的加速效果;

您可以在部署在线服务时在【高级设置】这里配置相应环境变量来开启对应功能:

- QUANTIZATION: 量化模式, 我们为 Hunyuan-Large 模型默认开启了 "ifq_nf4" 量化模式, 以适配更多机型部署。您也可以手动指定量化模式, 例如 NVIDIA Hopper 系列机型使用 "fp8" 量化模式;
- USE_LOOKAHEAD: 默认为 "0", 设置为 "1" 表示开启 Lookahead 并行解码;
- NUM_SPECULATIVE_TOKENS: 默认为 "6", 表示 Lookahead 并行解码一次解码长度, 若实际需要支持的并发数较大, 可以调小此值, 并发数小, 可以调大此值;
- MAX_MODEL_LEN: 模型上下文长度, 我们默认调小了此参数到最大 8192, 以适配更多机型部署, 您可以手动修改此值;

LLM 大模型系列

快速部署和体验内置开源 LLM 大模型

最近更新时间：2025-02-04 10:53:11

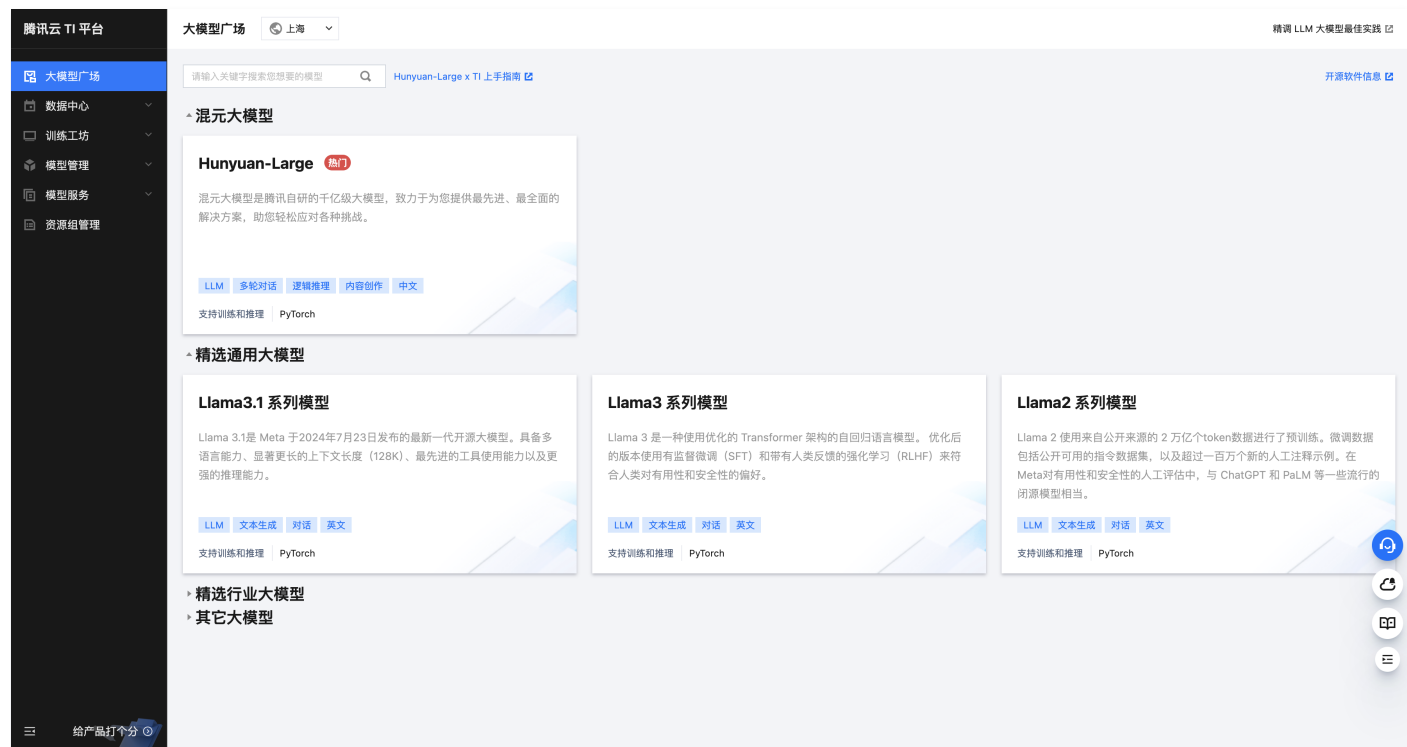
腾讯云 TI 平台 已全面接入混元、Llama3、Llama2 等主流开源大模型。

本文档将为您介绍如何通过 TI 平台，一键部署开源大模型，快速试一试模型推理效果，助力用户快速选型适合您业务场景的开源大模型。

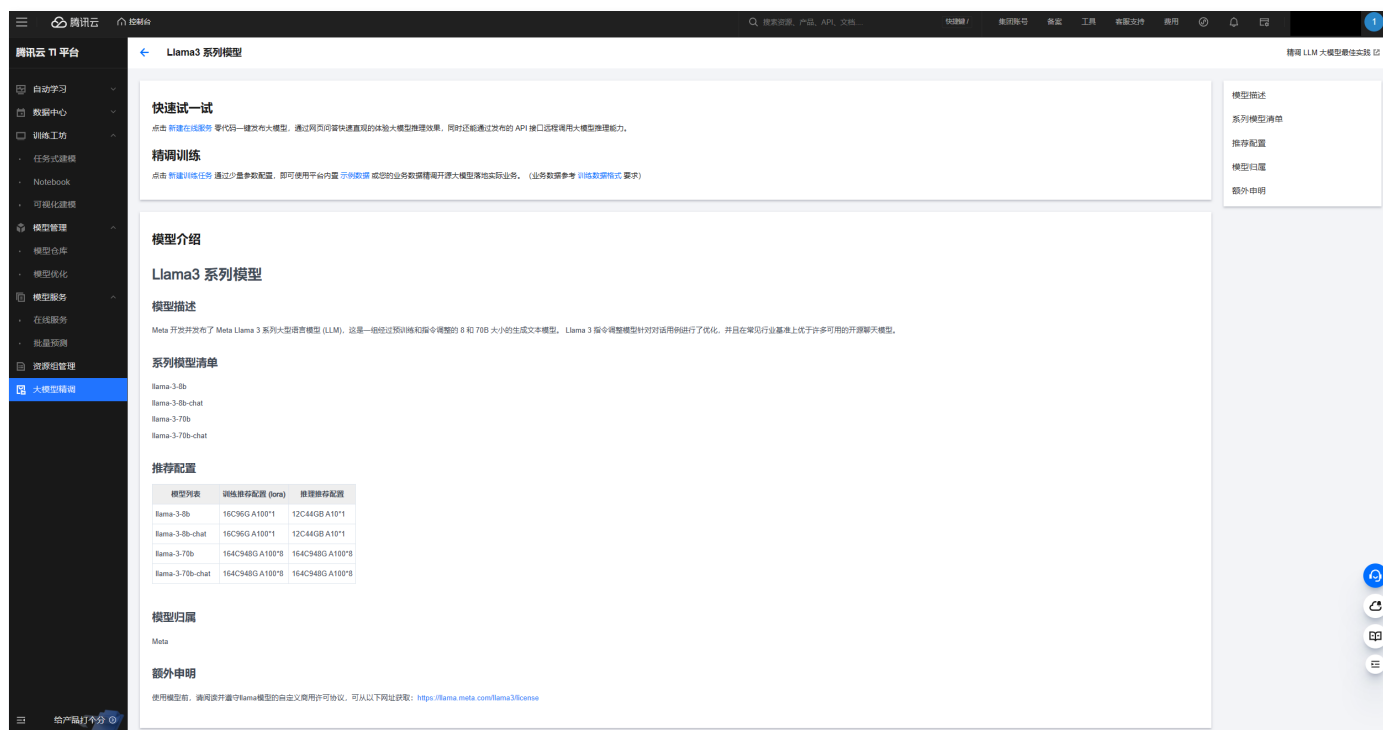
TI 平台已最新上架 DeepSeek 系列模型。如需进行部署和体验，可参考 [快速部署和体验 DeepSeek 系列模型](#)。

第一步：预览 TI 内置大模型清单

登录 [腾讯云 TI 平台](#)，在大模型广场首页，您可看到 TI 内置上架的开源大模型清单列表。



单击模型卡片，可进入每个模型页面查看详细介绍。



第二步：一键部署开源大模型

在模型详情页面，您可单击**新建在线服务**按钮，启动部署该大模型。本文档以部署“Llama3系列模型”中的llama_3_8b_chat模型为例。



单击后，平台会自动为您跳转到**模型服务 > 在线服务**的创建服务参数配置页面。

腾讯云 TI 平台

自动学习

数据中心

训练工坊

任务式建模

Notebook

可视化建模

模型管理

模型仓库

模型优化

模型服务

在线服务

批量预测

资源组管理

大模型精调

创建服务

服务名称 *

llama3-demo

请输入不超过60个字符，仅支持中英文、数字、下划线“_”、短横线“-”，只能以中英文、数字开头

服务版本

V1

服务描述 (选填)

请输入不超过500个字

0 / 500

地域 *

上海

计费模式 *

☒ 按量计费 ☐ 包年包月 (资源组)

服务实例 *

模型来源

☐ 模型仓库 ☐ CFS ☒ 镜像

模型和运行环境 *

内置大模型 / Llama3 系列模型 / llama_3_8b_chat

算力规格 *

12C44GB A10*1

高级设置

请求限流 ①

☒ 不限流 ☐ 单实例QPS ☐ 单实例最大并发数

实例调节 ①

☒ 手动调节 ☐ 自动调节

实例数量 *

- 1 +

是否生成鉴权

☐ 开启鉴权后，服务调用时会进行签名认证，已启动的服务可在服务调用页面查看签名密钥及签名计算指引

CLS 日志投递 ①

☐ TI 控制台会默认展示 7 天的日志，若您期望持久化存储日志，获得日志检索等服务，可以开启 CLS 日志投递，CLS 产品介绍和收集指南请查看[文档](#)

服务访问云上VPC资源 ①

☐ 开启后配置VPC和子网，会在您账户下的该子网，创建一个弹性网卡进行绑定，独立计费，弹性网卡当前未计费

自动停止

☐ 开启后，在线服务将在指定的停止时间自动停止，同时停止服务算力计费

标签 ①

标签键 标签值 ×

+ 添加 键值粘贴板

配置价格

服务配置费用 ① 6.26 元/小时

☐ 遵守平台要求，授权并同意 [《腾讯云 TI-ONE 训练平台服务协议》](#)

启动服务 取消

本页面重点参数配置如下：（其余参数都可使用默认值，不修改或按需修改）

- 服务名称：输入您的自定义服务名称。
- 计费模式：部署大模型推理服务支持“按量计费”和“包年包月”两种模式都可以。本示例采用“按量计费”。
- 服务实例
 - 模型来源：选择“镜像”类型。
 - 模型和运行环境：系统已为您内置默认填充了“内置大模型/Llama3系列模型/llama_3_8b_chat”，也可选择其他内置模型，如Llama3系列下的 llama_3_8b、llama_3_70b_chat、llama_3_70b等。
 - 算力规格：本示例选择 12C44GB A10*1 的机器配置，此外还支持V100、A100等机型。
详情请参见 [大模型推理所需资源指南](#)。
- 单击同意开源大模型的相关法律协议后，即可单击启动服务按钮。

第三步：在线体验模型效果

创建服务后，前端自动返回[模型服务 > 在线服务](#)列表页面，待大模型服务状态为“运行中”中后，表示服务已成功启动运行。

在线服务 上海

在线服务简介

新增服务

请输入名称或ID

名称	状态	计费模式	运行中总副本数	运行中 Pod 数量	标签	创建时间	操作
llama3-demo	运行中	按量计费	1/1	1/1	-	2024-05-13 14:58:41	调用API 查看日志 在线体验 删除

服务调用分为两种方式：

1. API 调用

您可以通过单击操作 > 调用API 按钮进入模型调用指导文档页面。文档中建议您通过API Explorer 配置调用代码。其中，需填写到 API Explorer 的“Model”参数取值需参考在线服务>名称>服务管理页面展示的“ID”参数。

llama3-demo

在线服务简介

服务管理 在线体验

名称	ID	服务版本	状态	计费模式	运行中 Pod 数量	服务详情	流量分配 (%)	当前占用资源	描述	创建时间	操作
llama3-demo	ms-Q0lgi46-1	V1	运行中	按量计费	1/1	山		12C44G A10*1	-	2024-05-13 14:58:41	停止 删除

2. 在线体验

您可以通过单击操作 > 在线体验 按钮进入模型快速体验页面。您可通过前端页面问答形式测试精调后的模型效果。

腾讯云 TI 平台

llama3-demo

服务管理 在线体验

自动学习
数据中心
训练工具
任务式建模
Notebook
可视化训练
模型管理
模型仓库
模型优化
模型服务
在线服务
批量训练
资源管理
大模型编辑

请输入您的问题，Enter 提交

重要提示 添加

第四步：管理模型服务，查看更多详情信息

您还能通过模型服务 > 在线服务列表页面单击对应服务名称进入详细的服务管理页面。

llama3-demo

在线服务简介

服务管理 在线体验

名称	ID	服务版本	状态	计费模式	运行中 Pod 数量	服务详情	流量分配 (%)	当前占用资源	描述	创建时间	操作
llama3-demo	ms-Q0lgi46-1	V1	运行中	按量计费	1/1	山		12C44G A10*1	-	2024-05-13 14:58:41	停止 删除

您可通过服务管理页面的操作 > 停止/重启按钮控制服务状态，并通过单击“服务详情”的 icon 查看监控/日志；也可以继续单击服务名称查看更详细的基本信息/实例列表/监控/事件/日志/更新记录等信息。

基本信息

服务ID	ms-f26tgk46-1
服务版本	V1
服务组ID	ms-f26tgk46
创建时间	2024-05-13 14:58:41
计费模式	按量计费
描述	-
模型来源	镜像
模型文件	内置大模型: Llama3 系列模型-llama_3_8b_chat
运行环境	tione.tencentcloudcr.com/qcloud-ti-platform/llm-infer:angel-vllm-0.3.3-ti-1.0.9
模型推理文件	-
模型热更新	未开启
单实例占用资源	12C44G A10*1
请求限流	未开启 
实例调节策略	手动调节
实例数量 1 个	
是否开启TI-ONE	未开启
内网访问外部	
自动停止	未开启
标签	-

导入和部署自定义 LLM 大模型（平台内置推理镜像）

最近更新时间：2025-06-12 14:27:11

总览

本文以【Qwen2-7B-Instruct】模型为例，指导如何将自定义大模型导入到 TI 平台，并使用平台内置推理镜像部署大模型对话推理服务。

前置要求

申请 CFS 或 GooseFSx

在导入和部署自定义 LLM 大模型中，您的大模型文件使用到的存储可以为 CFS 或者 GooseFSx，所以需要您首先申请 CFS 或者 GooseFSx，详情请查看[文件存储-创建文件系统及挂载点](#)或者[数据加速器 GooseFS](#)。

操作步骤

1. 上传模型文件到 CFS 或 GooseFSx

登录腾讯云 TI-ONE 控制台[训练工坊](#) > **Notebook**，单击**新建**，其中各字段的填写说明如下：

- 镜像**：选择任意**内置镜像**即可。
- 计费模式**：选择按量计费或包年包月均可，平台支持的计费规则请您查看[计费概述](#)。
- 存储配置**：选择 CFS 或 GooseFSx 文件系统，路径默认为根目录 /，用于指定保存用户自定义大模型位置。
- 其它设置**：默认不需要填写。

说明：
本 Notebook 实例仅用于上传或下载大模型文件。

腾讯云 TI 平台

自动学习

数据中心

训练工坊

任务式建模

Notebook

可视化建模

模型管理

模型服务

资源组管理

大模型精调

新建Notebook

名称 *

Qwen2-test

请输入不超过60个字符，仅支持中英文、数字、下划线"_"、短横"-"，只能以中英文、数字开头

地域 *

上海

镜像 *

内置镜像 / tilearn-llm1.0-torch2.1-angel-vllm1.0-py3.10-cuda12.1-gpu

计费模式 *

按量计费

包年包月

算力规格 *

28C116GB A10*1

存储配置 *

若要挂载多个CFS或CFS Turbo，请确保选择的文件系统在同一个VPC或同一个云联网下，系统会默认打通选择的第一个文件系统所在的VPC网络或云联网网络，其中云硬盘仅支持挂载一个实例；配置的第一个容器挂载路径会做为默认工作目录

数据类型	CFS文件系统	源路径 ①	容器挂载路径 ①	操作
CFS	test5875cmount (cfs-o8bth0up)	/	/home/tione/notebook	删除

+ 添加

标签 ①

标签键

标签值

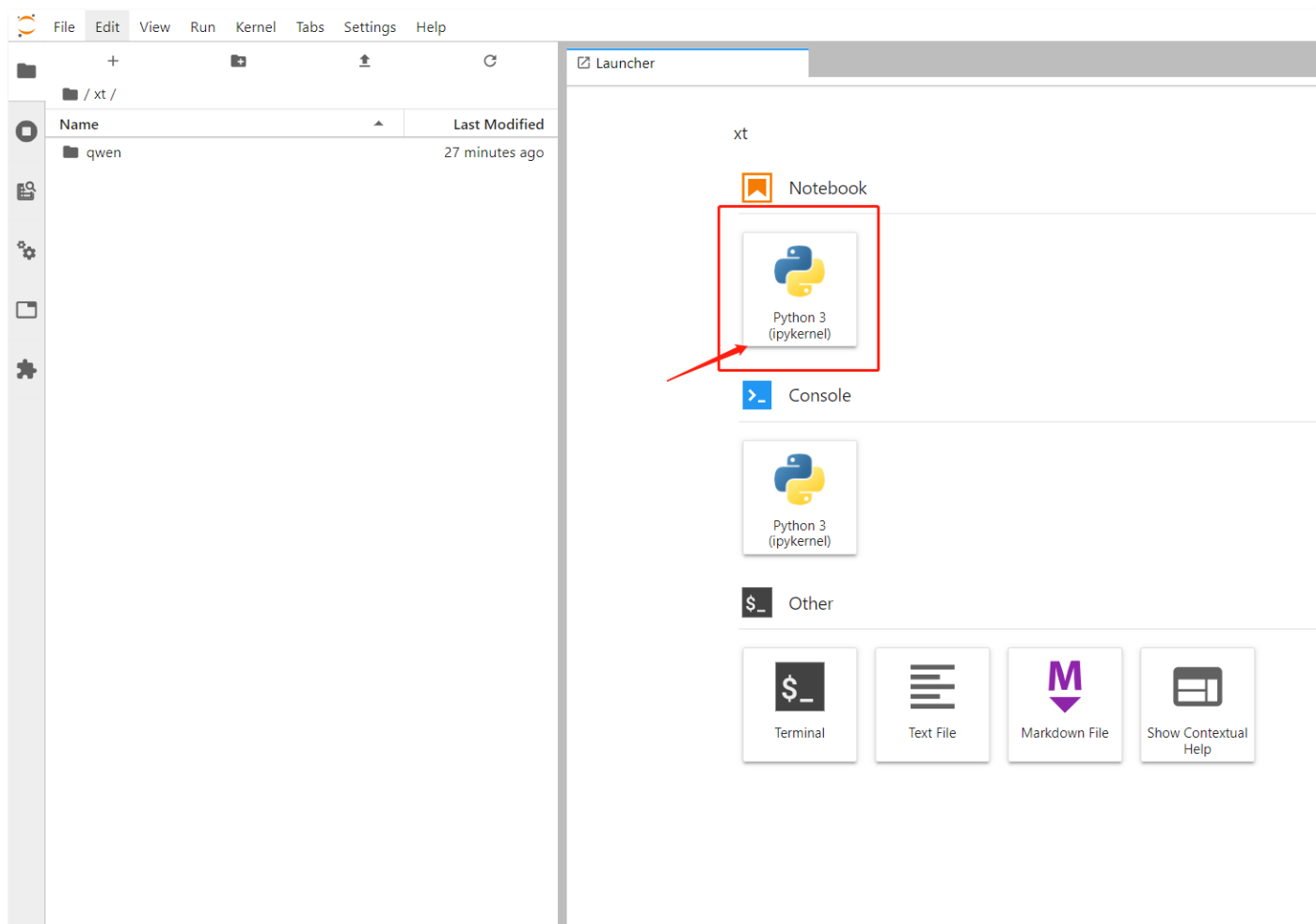
x

+ 添加

键值粘贴板

高级设置

新建成功后启动 Notebook，单击 **Notebook** > **Python3(ipynkernel)** 通过脚本下载所需模型；



您可在 [魔搭社区](#) 或 [Hugging Face](#) 检索需要用到的大模型，通过社区中 Python 脚本自行下载模型并保存到CFS中，本文以【Qwen2-7B-Instruct】模型为例，下载代码如下：

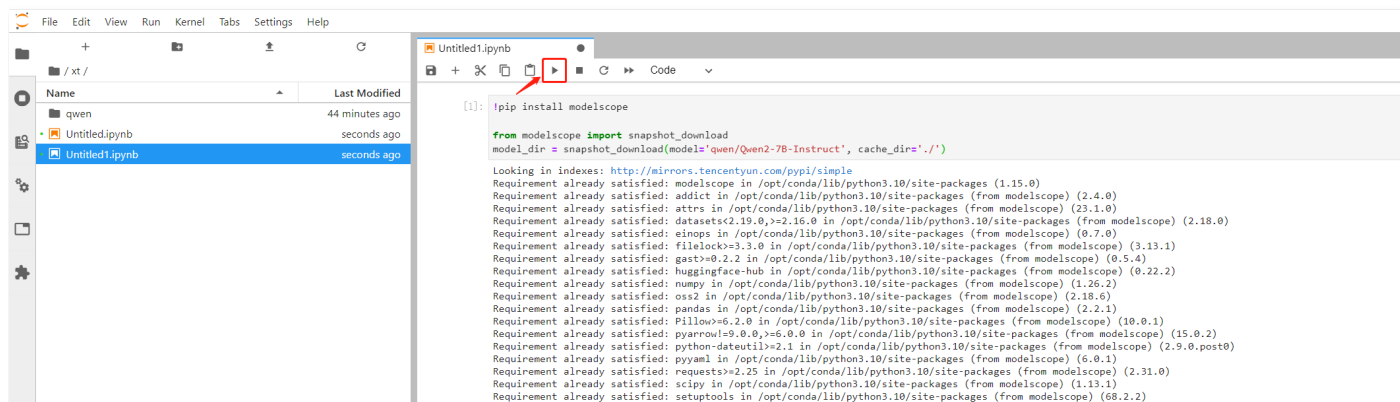
```
!pip install modelscope

from modelscope import snapshot_download
#qwen/Qwen2-7B-Instruct为需要下载模型名称，cache_dir为下载模型保存的地址，这里'./'表示将下载模型保存在CFS的根目录中
model_dir = snapshot_download('qwen/Qwen2-7B-Instruct', cache_dir='./')
```

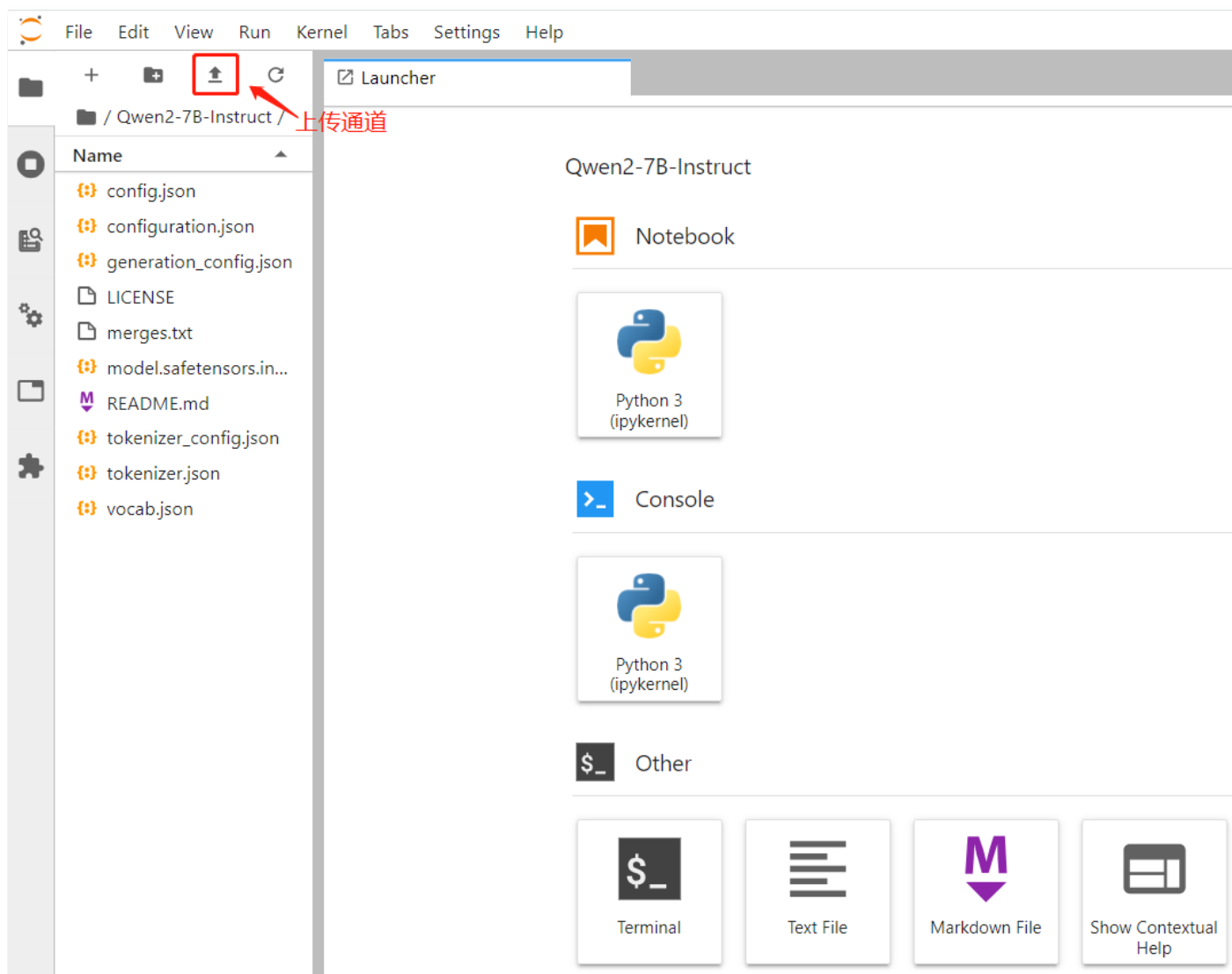
说明：

指定下载模型的地址 cache_dir（例如path/to/local/dir）后，后续在线服务 CFS 中指定模型地址为 /path/to/local/dir/qwen/Qwen2-7B-Instruct。

复制上述下载脚本并更换需要下载的模型后，粘贴到新建的 ipynb 文件中，点击运行按钮即可开始下载模型；



此外您也可以在本地下载或微调后，通过 notebook 上传通道将模型文件保存至 CFS 中，上传接口如图所示：



2. 创建在线服务

通过腾讯云TI平台的模型服务 > 在线服务，单击新建服务来启动推理服务，以下是服务实例配置的指引。

- **模型来源：**选择 CFS。
- **选择模型：**指定申请的 CFS，模型路径为 CFS 中下载或上传的模型路径，此处为【/qwen/Qwen2-7B-Instruct】。
- **运行环境：**选择【内置 / LLM / angel-vllm】。
- **算力规格：**根据实际的模型大小或拥有的资源情况选择，大模型推理时需要的机器资源与模型的参数量相关，推荐按如下规则配置推理服务资源。

模型参数量	GPU 卡类型和数量
-------	------------

6 ~ 8B	L20 * 1 / A10 * 1 / A100 * 1 / V100 * 1
12 ~ 14B	L20 * 1 / A10 * 2 / A100 * 1 / V100 * 2
65 ~ 72B	L20 * 8 / A100 * 8

- 【高级设置 > 环境变量】：需要设置模型名称MODEL_ID(魔搭社区或Hugging Face 上开源模型 ID)，以及对话模板名称CONV_TEMPLATE (若MODEL_ID与开源模型相同，可以不添加CONV_TEMPLATE参数)，常用的对话模板名称如下表所示，本文使用qwen-chat 系列，故设置为qwen-7b-chat。

对话模板名称CONV_TEMPLATE	支持的模型系列MODEL_ID
generate	非对话模型（直接生成，无对话模板）
llama-3	llama-3-8b-instruct、llama-3-70b-instruct 模型
llama-2	llama-2-chat 系列模型
qwen-7b-chat	qwen-chat 系列模型（chatml格式）
baichuan2-chat	baichuan2-chat 系列模型
baichuan-chat	baichuan-13b-chat 模型
chatglm3	chatglm3-6b 模型
chatglm2	chatglm2-6b 模型

在线服务创建实例配置如下：

模型来源

模型仓库

CFS

镜像

选择模型 ⓘ *

test5875cmount (cfs-o8bth0up) ▼ ↺

CFS控制台

/Qwen2-7B-Instruct

运行环境 *

内置 / LLM(1.0.0) / angel-vllm ▼

算力规格 *

28C116GB A10*1 ▼

▲ 高级设置

启动命令 ⓘ

请输入启动命令

环境变量

MODEL_ID

Qwen2-7B-Instruct

×

CONV_TEMPLATE

qwen-7b-chat

×

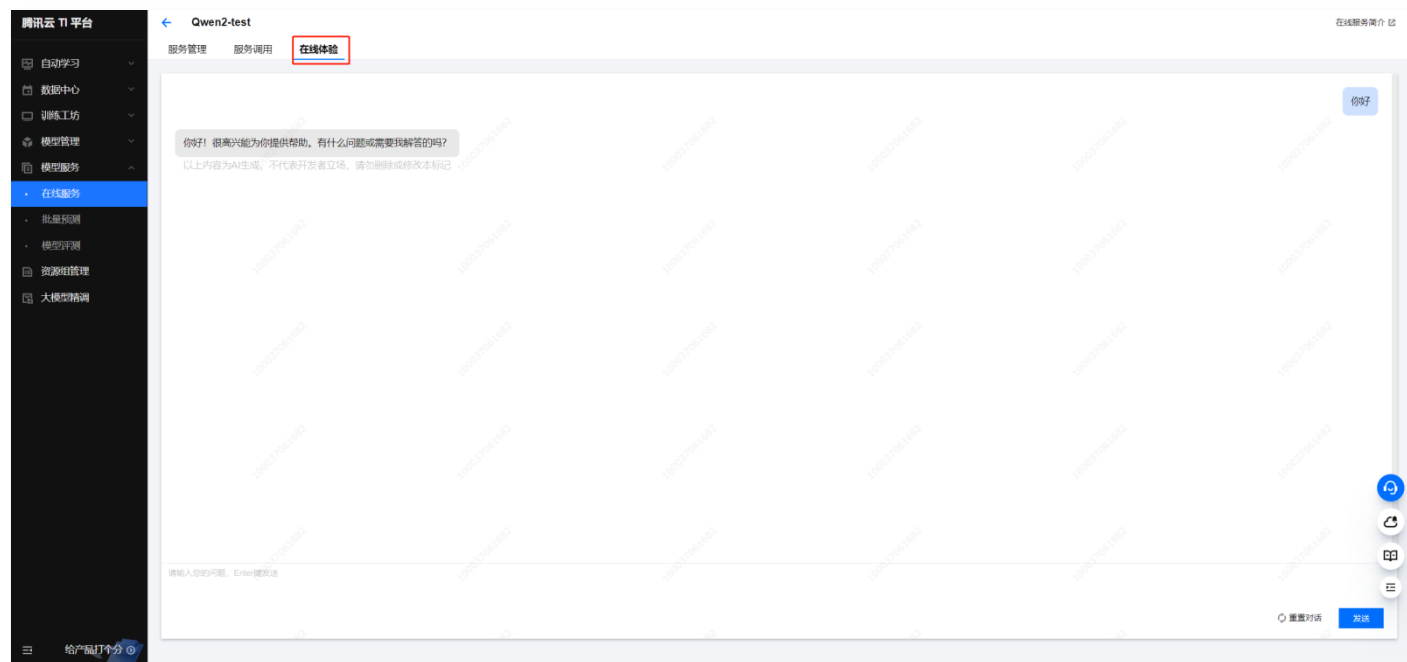
+ 新增变量

- ⚠ 注意：
- 如果对推理速度有较高要求，推荐您开启量化加速，通过环境变量 QUANTIZATION 设置，可选值有"none", "ifq", "smoothquant", "auto"。
 - none：表示关闭量化加速
 - ifq：表示开启在线 Int8 Weight-Only 量化，可以在效果基本不损失的情况下加速推理，并减少模型权重的显存占用。
 - smoothquant：表示开启 LayerwiseSearchSMQ 量化，可以在效果略微损失的情况下进一步加速推理（依赖提前准备量化后的模型文件，当前仅部分模型支持）
 - auto：表示自动判断量化模式，其中：
 - 若机型的显卡不支持量化，自动关闭量化。
 - 若模型目录中包含 smoothq_model-8bit-auto.safetensors 文件，会自动开启 LayerwiseSearchSMQ 量化加速。
 - 其他情况下，默认开启在线 Int8 Weight-Only 量化加速(ifq)。

- 若开启服务后日志报错 **CUDA out of memory**，此处由于模型 `max-model-len` 参数默认值 32k 较大（推理服务支持的最大上下文 token 数，默认为自动读取模型配置信息的上下文长度，若模型加载默认的上下文长度较大会导致显存不足），可通过环境变量 `MAX_MODEL_LEN` 来设置较小的数值（例如 16k 或 8k），也可以通过开启量化加速减少模型权重所占用的显存。

3. 前端在线体验

进入创建的在线服务详情，通过点击**在线体验** Tab 页即可与部署的大模型进行交互体验。



4. 接口服务调用

可通过**服务调用** Tab 页中的**接口信息** > **调用方式(在线测试)**进行访问，接口的调用地址为 `${SERVER_URL}/v1/chat/completions`，请求体的格式：

```
{ "messages": [ { "role": "user", "content": "你是谁" } ] }
```

字段 `content` 为具体的消息内容。

接口信息

接口调用地址: `https://ms-gp6j2-...gw.ap-shanghai.ti.tencentcs.com/ms-gp6j2/` `/v1/chat/completions`

服务类型: HTTP

请求方法: POST

调用方式(命令行): `curl -X POST https://ms-gp6j2-...w.ap-shanghai.ti.tencentcs.com/ms-gp6j2/v1/chat/completions -H 'Content-Type: application/json' -d '{ "messages": [ { "role": "user", "content": "你是谁" } ] }'`

若服务开启了鉴权，请参考[文档](#) [指引](#)调用

调用方式(在线测试): 请求体(Request Body 600KB 内)

```
1 { "messages": [ { "role": "user", "content": "你是谁" }, { "role": "assistant", "content": "你好" } ] }
```

发送请求

请求响应(Response)

```
1 Status: 200 OK
2 Connection: keep-alive
3 Content-Length: 355
4 Content-Type: application/json
5 Date: Tue, 25 Jun 2024 07:31:48 GMT
6 X-Ratelimit-Limit: 2000
7 X-Ratelimit-Remaining: 1999
8 X-Tigateway-Upstream-Status: 200
9
10 {
11   "id": "chatcmpl-x7MtnGm4i3cbk4w6GCT6YXR",
12   "object": "chat.completion",
13   "created": 1719300708,
14   "model": "Qwen2-7B-Instruct",
15   "choices": [
16     {
17       "index": 0,
18       "message": {
19         "role": "assistant",
20         "content": "我是阿里云开发的一款超大规模语言模型，我叫通义千问。"
21       }
22     }
23   ]
24 }
```

公网访问地址可从在线服务实例**服务调用**中获取，API 调用示例如下：

```
# 公网访问地址
```

```
SERVER_URL = https://ms-gp6rjk2jj-*****.gw.ap-shanghai.ti.tencentcs.com/ms-gp6rjk2j

# 非流式调用
curl -H "content-type: application/json" ${SERVER_URL}/v1/chat/completions -d '{"messages":[{"role":
"user", "content": "你好"}], "temperature": 0.0}'

# 流式调用
curl -H "content-type: application/json" ${SERVER_URL}/v1/chat/completions -d '{"messages":[{"role":
"user", "content": "你好"}], "temperature": 0.0, "stream": true}'
```

非流式返回结果：

```
{ "id": "chatcmpl-4aeRgYwnaYe4RzmmcyKtYs", "object": "chat.completion", "created": 1698291242, "model": "baichuan-
13b-chat", "choices": [{"index": 0, "message": {"role": "assistant", "content": "你好！有什么我能帮到你的
吗?"}, "finish_reason": "stop"}], "usage": {"prompt_tokens": 4, "total_tokens": 16, "completion_tokens": 12}}
```

流式返回结果：

```
data: {"id": "chatcmpl-hn5mCVt4szVVZBa4fVFZWF", "model": "baichuan-13b-chat", "choices": [{"index": 0,
"delta": {"role": "assistant"}, "finish_reason": null}]}

data: {"id": "chatcmpl-hn5mCVt4szVVZBa4fVFZWF", "model": "baichuan-13b-chat", "choices": [{"index": 0,
"delta": {"content": " 你"}, "finish_reason": null}], "usage": {"prompt_tokens": 4, "total_tokens": 5,
"completion_tokens": 1}}

data: {"id": "chatcmpl-hn5mCVt4szVVZBa4fVFZWF", "model": "baichuan-13b-chat", "choices": [{"index": 0,
"delta": {"content": "好"}, "finish_reason": null}], "usage": {"prompt_tokens": 4, "total_tokens": 6,
"completion_tokens": 2}}

.....此处省略中间结果.....

data: {"id": "chatcmpl-hn5mCVt4szVVZBa4fVFZWF", "model": "baichuan-13b-chat", "choices": [{"index": 0,
"delta": {"content": "?"}, "finish_reason": null}], "usage": {"prompt_tokens": 4, "total_tokens": 15,
"completion_tokens": 11}}

data: {"id": "chatcmpl-hn5mCVt4szVVZBa4fVFZWF", "object": "chat.completion.chunk", "created": 1714120317,
"model": "baichuan-13b-chat", "choices": [{"index": 0, "delta": {}, "finish_reason": "stop"}], "usage":
{"prompt_tokens": 4, "total_tokens": 16, "completion_tokens": 12}}

data: [DONE]
```

另外也可以通过 python 常用的 requests 库来使用服务，下面是一个命令行与 Qwen2-7B-Instruct 大模型推理服务进行对话交互的 Demo 示例：

```
import argparse
import requests
import json

def chat(messages):
    data = {
        "messages": messages,
        "temperature": args.temperature,
        "max_tokens": args.max_tokens,
        "top_p": args.top_p,
        "stream": True, # 开启流式输出
    }
    header = {
        "Content-Type": "application/json",
    }
    if args.token:
```

```
header["Authorization"] = f"Bearer {args.token}"

response = requests.post(f"{args.server}/v1/chat/completions", json=data, headers=header, stream=True)
# 设置 stream=True 参数获取实时数据流
if response.status_code != 200:
    print(response.json())
    exit()

result = ""
print("Assistant: ", end = "", flush = True)
for part in response.iter_lines():
    if part:
        if "content" in part.decode("utf-8"):
            content = json.loads(part.decode("utf-8")[5:])[0]["delta"]["content"] # 字符串
            # 过滤 data: 之后转成 json 格式再取文本
            result += content
            print(content, end = "", flush = True)
print("\n")

return result

if __name__ == "__main__":
    parser = argparse.ArgumentParser(description="Chat CLI Demo.")
    # --server 替换为实际的调用地址
    parser.add_argument("--server", type=str, default="https://ms-gp6rjk2j-*****.gw.ap-shanghai.ti.tencentcs.com/ms-gp6rjk2j")
    parser.add_argument("--max-tokens", type=int, default=512)
    parser.add_argument("--temperature", type=float, default=0.7)
    parser.add_argument("--top_p", type=float, default=1.0)
    parser.add_argument("--system", type=str, default=None)
    parser.add_argument("--token", type=str, default=None)
    args = parser.parse_args()

    messages = []
    if args.system:
        messages.append({"role": "system", "content": args.system})

    while True:
        user_input = input("User: ")
        messages.append({"role": "user", "content": user_input})
        response = chat(messages)
        messages.append({"role": "assistant", "content": response})
```

精调内置开源大模型

最近更新时间：2025-06-12 14:27:11

步骤速览

一、新建 CFS 或者 GooseFSx：在 CFS 或者 GooseFSx 控制台，新建自己的 CFS 或者 GooseFSx 文件存储系统。

二、挂载 CFS 或者 GooseFSx：新建并启动开发机，挂载 CFS 或者 GooseFSx 后管理 CFS 或者 GooseFSx 文件目录。（若您希望使用自己的业务数据精调大模型，则可在此步同时上传业务数据文件到 CFS 或者 GooseFSx；当然平台也为您内置了一份示例数据用于快速启动精调任务）。

三、快速试一试，选型基底模型：在“大模型广场”界面下，选择合适模型，快速试一试。若您后续采用内置示例数据精调模型，则可用以下问题测试精调前基底模型效果。

测试 prompt1 内容:

```
"Explain the text in one sentence. \n Text:(CNN) -- Actress Lindsay Lohan's father jumped from a third-floor balcony Thursday to try to escape re-arrest, police said, but officers soon caught up with him and took him into custody -- again.\n\nOfficers detained Michael Lohan after his girlfriend, Kate Major, said he violated terms of his release from jail earlier this week by calling her, the Tampa Police Department said in a statement. Investigators arrested Lohan on suspicion of domestic violence after an incident on Monday night, but they let him out of jail after he posted bail.\n\nJust after 1 a.m. Thursday, Lohan's girlfriend called police to say Lohan had violated the terms of his release by calling her, the police statement said. Lohan called his girlfriend again while she was talking with police, and she put him on speakerphone, it said.\n\nAfter prosecutors authorized Lohan's re-arrest, officers went to a hotel, the Tahitian Inn, where he was staying.\n\n"Upon seeing officers, he jumped out of a third-story balcony in an attempt to escape arrest," the police statement said.\n\nOfficers took him into custody after a short chase and discovered after taking him to jail that he may have broken his foot, police said. Doctors were evaluating him.\n\nLohan is not expected to be released from Tampa General Hospital Thursday and has been admitted for overnight observation, said Tampa police spokeswoman Andrea Davis.\n\nPolice responded to a domestic violence call at Michael Lohan's home Monday and arrested him for the battery of his live-in girlfriend, police said.\n\nLohan told reporters Wednesday he "didn't lay a hand" on his girlfriend.\n\n"I did not hurt her," he said.\n\nLohan also said he never was served with a restraining order to stay away from the woman. According to CNN Tampa affiliate WFTS, a Sarasota County judge issued a temporary restraining order Tuesday.\n\nLohan, 51, was arrested in Los Angeles in March in connection with alleged domestic violence and was charged with one misdemeanor count of corporal injury to a cohabitant.\n\nThe relationship between Lohan and his daughter has been publicly strained for years, although the two did undergo family counseling together during her treatment at the Betty Ford Center.\n\nCNN's Rich Phillips contributed to this report.\n\n"
```

测试 prompt2 内容:

```
"Explain the text in one sentence. \n Text:(CNN) -- A federal judge Friday ruled in favor of a former UCLA college basketball star who sued to end the NCAA's control over the rights to college athletes' names, images and likenesses.\n\nIn a landmark decision, U.S. District Judge Claudia Wilken sided with Ed O'Bannon in his lawsuit against the National Collegiate Athletic Association. O'Bannon argued athletes in the top tier of college basketball and football should be allowed to profit from their schools' use of their likenesses.\n\nIn a 99-page ruling, Wilken wrote that current NCAA rules "unreasonably restrain trade in the market for certain educational and athletic opportunities offered by NCAA Division I schools."\n\nWilken issued an injunction to block the NCAA from prohibiting its member schools and conferences from offering their Football Bowl Subdivision or Division I basketball recruits a limited share of the revenues generated from the use of their names, images, and likenesses.\n\nShe did rule, however, that the NCAA could set a cap on the money paid to athletes, as long as it allows at least $5,000 per athlete per year.\n\n"The NCAA's witnesses stated that their concerns about student-athlete compensation would be minimized or negated if compensation was capped at a few thousand dollars per year," the judge wrote.\n\nO'Bannon's suit alleged the waivers the athletes are required to sign are illegal and asked that players be able to collectively negotiate the terms of their likenesses in order to keep a share of those profits.\n\n"Before the court in this case is only whether the NCAA violates antitrust law by agreeing with its member schools to restrain their ability to compensate Division I men's basketball and FBS football players any more than the current association rules allow," Wilken wrote. "For the reasons set forth above, the court finds that this restraint does violate antitrust law."\n\nThe ruling could potentially change college sports
```

drastically, eventually forcing the NCAA to restructure its amateur model and allow college athletes to be paid.

NCAA chief legal officer Donald Remy said: "We disagree with the court's decision that NCAA rules violate antitrust laws. We note that the court's decision sets limits on compensation, but are reviewing the full decision and will provide further comment later."

William Isaacson, an attorney for the plaintiffs, called the ruling "a big step forward for common decency."

"One of the things the judge is saying here .. is some sharing is OK," he said. "It won't affect amateurism, won't affect the popularity of the sport. She made a very reasonable and significant and measured decision."

Ramogi Huma, president of the National College Players Association, said the ruling was "a big win" that signaled "the time for college athletes to get their due," though he was critical of the \$5,000 compensation cap.

"The ruling says the NCAA was operating illegally and college athletes do have rights," said Huma, who helped find players to join the case. "Even if you label them student-athletes and want to call it amateurism, it doesn't give the NCAA the right to deny them the rights that other Americans deserve."

Sonny Vaccaro, who started the lawsuit by introducing O'Bannon and lead attorney Michael Hausfeld, said the decision was precedent-setting.

"The key is, they're allowed to get paid," said Vaccaro, who helped pioneer branding athletes by putting Nike shoes on Michael Jordan.

Vaccaro called the decision his most important career accomplishment.

"It's more important to me that these kids won and this go forward and the principles are right," he said. "To me, it's more important than the Jordan and Kobe and things I did in my professional life. ... This was just something that was wrong and I totally believed in it."

The judge wrote that the injunction will not affect student-athletes who enroll in college before July 1, 2016.

When athletes commit to a university, players are required to sign a waiver that relinquishes their right to their own likenesses in every form.

That means they can't make money off their television appearances, their jerseys, or in any other way.

The universities get any revenues from selling sports paraphernalia or other material related to the players.

The trial began June 9 in federal court in Oakland, California. The plaintiffs were 20 current and former student athletes who play or played for an FBS football or Division I men's basketball team starting in 1956.

Legal appeals could delay a final outcome for years but the decision is in a position to be the first major NCAA reform effort to take hold.

Already the issues brought up in the case have had an effect, even before the ruling was made.

Texas A&M, the University of Arizona and Northwestern University have decided to stop selling jerseys with the numbers of specific players.

Instead, Texas A&M will sell the number 12 jersey, in keeping with its 12th man tradition; and Arizona will sell jerseys with numbers that correspond to the year of competition -- 14 for this year, according to a school spokesman. Northwestern will sell only jersey number 51, in honor of its head coach, Pat Fitzgerald, and legendary Chicago linebacker Dick Butkus.

The NCAA's argument in both the ongoing O'Bannon suit and another one filed by former quarterback Sam Keller, also in federal court in Oakland, is that it is trying to protect the amateur model of college sports.

Paying college athletes would hurt traditions, NCAA chief testifies

NCAA under fire: 5 things to know

NCAA faces change, legal challenges in months ahead

四、启动训练任务：在“大模型广场”界面下，选择合适模型，新建训练任务，完成相关配置后启动训练。重点关注以下配置：

- 训练资源：建议最低使用以下资源配置。

模型参数	CPU	GPU
7B	40C 800G	A100*8 或 A800*8
13B	120C 1800G	A800*8

- 训练输出路径：配置到第二步新建的 CFS 文件目录。
- 根据需要修改“调优参数”。

五、监控训练过程：在“任务式建模”界面下，单击“任务名称”，查看训练监控和日志。

六、查看训练输出文件：打开第二步的开发机，查看 CFS 或者 GooseFSx 上的训练输出文件。

七、部署模型服务：在“在线服务”界面下，新建服务，完成相关配置后，启动服务。重点关注以下配置：

- 推理资源：建议最低使用以下资源配置。

模型参数	CPU	GPU
7B	12C 44G	A10*1
13B	16C 96G	A100*1

八、服务调用：在“在线服务”界面下，调用部署好的服务。

- 在线体验：单击“在线体验”测试模型效果。可用第三步测试基底模型的同一批问题再次测试精调后的模型，对比训练前后效果。
- API 调用：通过 API 接口对接您的业务系统。

平台内置示例数据介绍：

英文 LLM 内置数据：1088 条英文新闻摘要生成数据对（[下载查看](#)）

中文 LLM 内置数据：1094 条中文新闻摘要生成数据对（[下载查看](#)）

前提条件

- 完成腾讯云 [TI-ONE 训练平台](#) 的服务注册与开通，请参考 [账号与权限说明](#)。
- 完成腾讯云 [私有网络](#) 服务注册与开通，并完成私有网络创建，请参考 [创建私有网络](#)。

详细流程

一、新建 CFS

CFS(Cloud File Storage, CFS) 文件存储系统官方介绍：[CFS 介绍文档](#)，CFS 产品 [控制台](#)。

数据加速器 GooseFS 文件系统官方介绍：[数据加速器 GooseFS](#)，产品控制台 [控制台](#)。

您新建的 CFS 或者 GooseFSx 文件存储系统在 TI 上的使用场景：

- 您用于精调大模型的业务数据需要上传到 CFS 或者 GooseFSx（您也可直接使用 TI 内置示例数据快速启动精调训练任务）。
- 大模型精调过程中，会产生一些模型中间结果（Checkpoint）、以及精调后最终得到的业务大模型等需要存储到您的 CFS。

二、挂载 CFS 或者 GooseFSx

本步骤的操作原因：

- 因为第一步中您新建的 CFS 下还只有一个根目录，您需要更好的组织 CFS 中的目录层级结构用于区分不同训练任务的不同输出路径，例如在根目录下新建一个“/sft-demo-20230825”的子目录作为本示例训练任务的输出路径。
- 若您希望用自己的业务数据精调大模型，则您也需要挂载 CFS 或者 GooseFSx 后，将训练数据文件上传到 CFS 或者 GooseFSx 中。

1. 新建开发机实例并挂载用户 CFS 或者 GooseFSx。

用户进入 [腾讯云 TI 平台](#)，在[训练工坊](#) > [开发机](#)二级页面，单击[新建](#)按钮，新建一个开发机实例，且将用户的 CFS 或者 GooseFSx 配置挂载到该容器实例上，参数如下，请重点关注，您此时新建的开发机所在地域需要和第一步中新建的 CFS 所在地域保持一致。

备注：

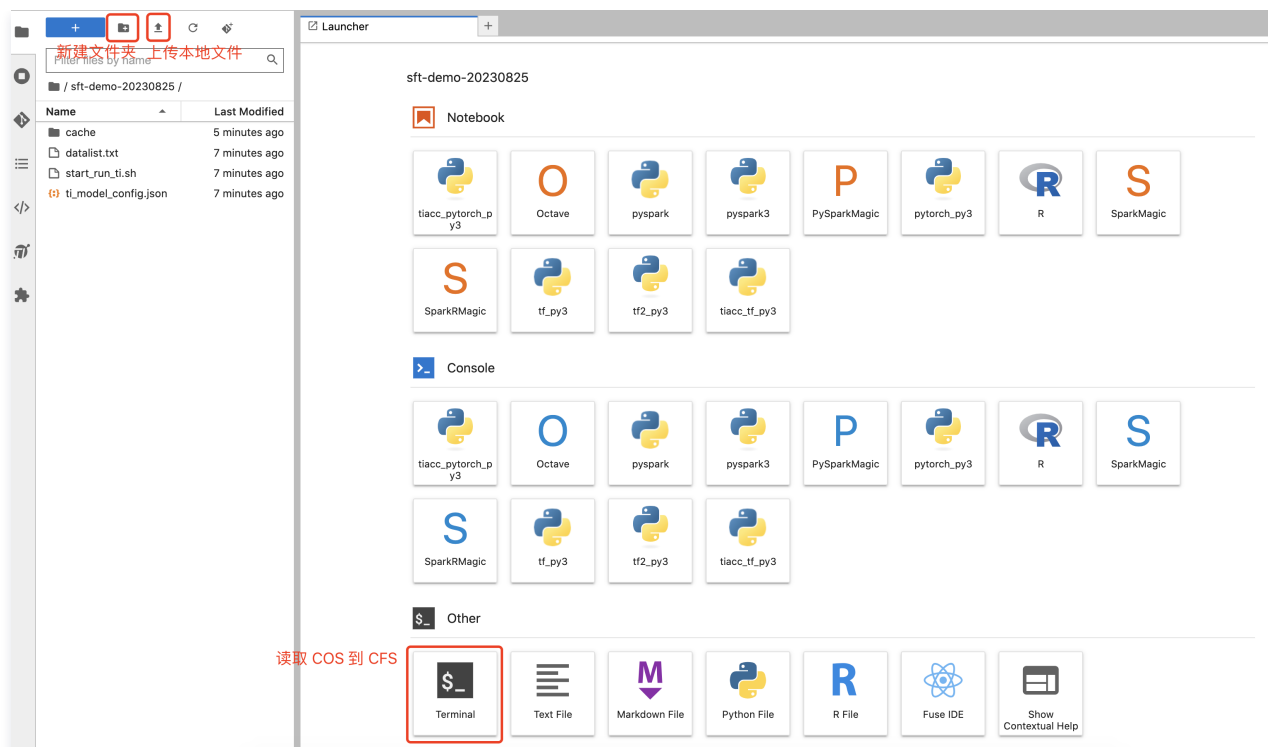
- 此时选择的“镜像”可在“内置镜像”中任意选择，因为该开发机仅用于挂载 CFS 或者 GooseFSx、可视化的管理 CFS 或者 GooseFSx 上的文件信息，对镜像中安装的依赖包没有特殊的要求。
- 此时选择的“算力规格”可用较小的“2C4G”启动开发机，因为这个资源仅仅用于启动开发机实例而不是最终大模型精调所用的训练资源。

2. 进入运行中的开发机

等待您新建的开发机实例状态处于“运行中”后，单击[操作](#) > [打开](#)按钮，进入开发机页面。

此时，您就可以通过开发机界面，直接操作您的 CFS 或者 GooseFSx 文件系统了。您有 3 种操作场景：

- 新建文件夹：管理 CFS 或者 GooseFSx 上的文件目录，如新建一个“sft-demo-20230825”的文件夹后续用于作为本示例训练任务的输出目录。
- 上传本地文件到 CFS 或者 GooseFSx：若您希望用自己存储在本地的业务数据精调模型，则您可在此界面上上传本地数据文件到 CFS。当然您也可直接使用平台为您内置的示例数据作为训练数据。
- 拉取 COS 数据到 CFS 或者 GooseFSx：打开“Terminal”，执行 coscmd 的相关命令，实现将存储在 COS 中的业务数据文件拉取到 CFS 文件系统中。可参考 [COSCMD 工具](#) 产品文档。



三、快速试一试，选型基底模型

1. 一键部署基底模型

用户进入 [腾讯云 TI 平台](#)，在**大模型广场**二级页面，选择您希望快速试一试模型效果的卡片，单击进入卡片详情。本示例以试一试“Llama2系列模型”的 llama_2_7b_chat模型为例，单击**新建在线服务**按钮，启动服务配置页面。



单击后，平台会自动为您跳转到**模型服务 > 在线服务**的创建服务参数配置页面。

← 创建服务

服务名称 *

请输入服务名称

请输入不超过60个字符，仅支持中英文、数字、下划线"_"、短横"-", 只能以中英文、数字开头

服务版本

V1

服务描述 (选填)

请输入不超过500个字

0 / 500

地域 *

上海

部署方式 *

☒ 标准部署 ①

☐ 多机分布式部署 ①

☐ 多角色部署(PD分离) ①

服务新建后，更新服务、新增版本均无法修改部署方式

机器来源 *

从CVM机器中选择

从第三方服务器中选择

从T1ONE平台购买

TI-ONE 收取软件和机器费用。详见 [产品定价](#)

计费模式 *

☒ 按量计费

☐ 包年包月

副本设置 *

请确保您选择的存储实例 (CFS) 和纳管资源组的节点网络互通

模型来源

☐ 模型仓库

☐ CFS

☒ 镜像

模型和运行环境 *

内置大模型 / Llama2 系列模型 / llama_2_7b_chat

开启gRPC ①

☐

端口 ① *

8501

算力规格 *

12C44GB A10*1

大模型推理所需资源指南

高级设置

配置价格

本页面重点参数配置如下：（其余参数都可使用默认值，不修改或按需修改）

- 服务名称：输入您的自定义服务名称。
- 计费模式：部署大模型推理服务支持“按量计费”和“包年包月”两种模式都可以。本示例采用“按量计费”。
- 服务实例
 - 模型来源：选择“镜像”类型。
 - 模型和运行环境：系统已为您内置默认填充了“内置大模型/Llama2系列模型/llama_2_7b_chat”，也可选择其他内置模型，如Llama2系列下的llama_2_7b、llama_2_13b_chat、llama_2_13b、llama_2_70b_chat、llama_2_70b等。
 - 算力规格：本示例选择 12C44GB A10*1 的机器配置，此外还支持V100、A100等机型（裸金属机型不支持）。
 - 单击同意开源大模型的相关法律协议后，即可单击启动服务按钮。

2. 服务调用

创建服务后，前端自动返回模型服务 > 在线服务列表页面，待大模型服务状态为“运行中”中后，表示服务已成功启动运行。

服务调用分为两种方式：

2.1 API 调用

您可以通过单击操作 > 调用API 按钮进入模型调用指导文档页面。文档中建议您通过 API Explorer 配置调用代码。

其中，需填写到 API Explorer 的“Model”参数取值需参考在线服务>名称>服务管理页面展示的“ID”参数。

名称	ID	服务版本	状态	机器来源	运行/部署副本数量	服务详情	流量分配 (%)	当前占用资源	描述	创建时间	操作
	ms-gsvm7n5-1	V1	已停止		0/0					2025-05-15 01:42...	启动 更新 删除 扩容

2.2 在线体验

您可以通过单击操作 > 在线体验 按钮进入模型快速体验页面。您可通过前端页面问答形式测试精调后的模型效果。

若您后续使用平台内置示例数据进行精调训练，则可尝试以下问题体验精调前基底模型效果。

测试 prompt1 内容:

```
"Explain the text in one sentence. \n Text:(CNN) -- Actress Lindsay Lohan's father jumped from a third-floor balcony Thursday to try to escape re-arrest, police said, but officers soon caught up with him and took him into custody -- again.\n\nOfficers detained Michael Lohan after his girlfriend, Kate Major, said he violated terms of his release from jail earlier this week by calling her, the Tampa Police Department said in a statement. Investigators arrested Lohan on suspicion of domestic violence after an incident on Monday night, but they let him out of jail after he posted bail.\n\nJust after 1 a.m. Thursday, Lohan's girlfriend called police to say Lohan had violated the terms of his release by calling her, the police statement said. Lohan called his girlfriend again while she was talking with police, and she put him on speakerphone, it said.\n\nAfter prosecutors authorized Lohan's re-arrest, officers went to a hotel, the Tahitian Inn, where he was staying.\n\nUpon seeing officers, he jumped out of a third-story balcony in an attempt to escape arrest,\" the police statement said.\n\nOfficers took him into custody after a short chase and discovered after taking him to jail that he may have broken his foot, police said. Doctors were evaluating him.\n\nLohan is not expected to be released from Tampa General Hospital Thursday and has been admitted for overnight observation, said Tampa police spokeswoman Andrea Davis.\n\nPolice responded to a domestic violence call at Michael Lohan's home Monday and arrested him for the battery of his live-in girlfriend, police said.\n\nLohan told reporters Wednesday he \"didn't lay a hand\" on his girlfriend.\n\n\"I did not hurt her,\" he said.\n\nLohan also said he never was served with a restraining order to stay away from the woman. According to CNN Tampa affiliate WFTS, a Sarasota County judge issued a temporary restraining order Tuesday.\n\nLohan, 51, was arrested in Los Angeles in March in connection with alleged domestic violence and was charged with one misdemeanor count of corporal injury to a cohabitant.\n\nThe relationship between Lohan and his daughter has been publicly strained for years, although the two did undergo family counseling together during her treatment at the Betty Ford Center.\n\nCNN's Rich Phillips contributed to this report.\n\n"
```

测试 prompt2 内容:

```
"Explain the text in one sentence. \n Text:(CNN) -- A federal judge Friday ruled in favor of a former UCLA college basketball star who sued to end the NCAA's control over the rights to college athletes' names, images and likenesses.\n\nIn a landmark decision, U.S. District Judge Claudia Wilken sided with Ed O'Bannon in his lawsuit against the National Collegiate Athletic Association. O'Bannon argued athletes in the top tier of college basketball and football should be allowed to profit from their schools' use of their likenesses.\n\nIn a 99-page ruling, Wilken wrote that current NCAA rules \"unreasonably restrain trade in the market for certain educational and athletic opportunities offered by NCAA Division I schools.\" \n\nWilken issued an injunction to block the NCAA from prohibiting its member schools and conferences from offering their Football Bowl Subdivision or Division I basketball recruits a limited share of the revenues generated from the use of their names, images, and likenesses.\n\nShe did rule, however, that the NCAA could set a cap on the money paid to athletes, as long as it allows at least $5,000 per athlete per year.\n\nThe NCAA's witnesses stated that their concerns about student-athlete compensation would be minimized or negated if compensation was capped at a few thousand dollars per year,\" the judge wrote.\n\nO'Bannon's suit alleged the waivers the athletes are required to sign are illegal and asked that players be able to collectively negotiate the terms of their likenesses in order to keep a share of those profits.\n\nBefore the court in this case is only whether the NCAA violates antitrust law by agreeing with its member schools to restrain their ability to compensate Division I men's basketball and FBS football players any more than the current association rules allow,\" Wilken wrote. \n\nFor the reasons set forth above, the court finds that this restraint does violate antitrust law.\n\nThe ruling could potentially change college sports drastically, eventually forcing the NCAA to restructure its amateur model and allow college athletes to be paid.\n\nNCAA chief legal officer Donald Remy said: \"We disagree with the court's decision that NCAA rules violate antitrust laws. We note that the court's decision sets limits on compensation, but are reviewing the full decision and will provide further comment later.\" \n\nWilliam Isaacson, an attorney for the plaintiffs, called the ruling \"a big step forward for common decency.\" \n\nOne of the things the judge is saying here .. is some sharing is OK,\" he said. \n\nIt won't affect amateurism, won't affect the popularity of the sport. She made a very reasonable and significant and measured decision.\" \n\nRamogi Huma, president of the National College Players Association, said the ruling was
```

```
\n"a big win\n" that signaled \n"the time for college athletes to get their due,\n" though he was critical of the $5,000 compensation cap.\n\n\n"The ruling says the NCAA was operating illegally and college athletes do have rights,\n" said Huma, who helped find players to join the case. \n"Even if you label them student-athletes and want to call it amateurism, it doesn't give the NCAA the right to deny them the rights that other Americans deserve.\n\n\nSonny Vaccaro, who started the lawsuit by introducing O'Bannon and lead attorney Michael Hausfeld, said the decision was precedent-setting.\n\n\n"The key is, they're allowed to get paid,\n" said Vaccaro, who helped pioneer branding athletes by putting Nike shoes on Michael Jordan.\n\n\nVaccaro called the decision his most important career accomplishment.\n\n\n"It's more important to me that these kids won and this go forward and the principles are right,\n" he said. \n"To me, it's more important than the Jordan and Kobe and things I did in my professional life. ...This was just something that was wrong and I totally believed in it.\n\n\nThe judge wrote that the injunction will not affect student-athletes who enroll in college before July 1, 2016.\n\n\nWhen athletes commit to a university, players are required to sign a waiver that relinquishes their right to their own likenesses in every form.\n\n\nThat means they can't make money off their television appearances, their jerseys, or in any other way.\n\n\nThe universities get any revenues from selling sports paraphernalia or other material related to the players.\n\n\nThe trial began June 9 in federal court in Oakland, California. The plaintiffs were 20 current and former student athletes who play or played for an FBS football or Division I men's basketball team starting in 1956.\n\n\nLegal appeals could delay a final outcome for years but the decision is in a position to be the first major NCAA reform effort to take hold.\n\n\nAlready the issues brought up in the case have had an effect, even before the ruling was made.\n\n\nTexas A&M, the University of Arizona and Northwestern University have decided to stop selling jerseys with the numbers of specific players.\n\n\nInstead, Texas A&M will sell the number 12 jersey, in keeping with its 12th man tradition; and Arizona will sell jerseys with numbers that correspond to the year of competition -- 14 for this year, according to a school spokesman. Northwestern will sell only jersey number 51, in honor of its head coach, Pat Fitzgerald, and legendary Chicago linebacker Dick Butkus.\n\n\nThe NCAA's argument in both the ongoing O'Bannon suit and another one filed by former quarterback Sam Keller, also in federal court in Oakland, is that it is trying to protect the amateur model of college sports.\n\n\nPaying college athletes would hurt traditions, NCAA chief testifies\n\n\nNCAA under fire: 5 things to know\n\n\nNCAA faces change, legal challenges in months ahead\n\n\n"
```

四、启动训练任务

1. 训练入口

用户进入 [腾讯云 TI 平台](#)，在**大模型广场**二级页面，选择您需要精调的大模型卡片，单击卡片进入模型详情页面。本示例以精调“Llama2系列模型”的 llama_2_7b_chat模型为例。

单击详情页面的“新建训练任务”即可开始零代码精调大模型。

其中，精调任务的训练数据来源有两种：

- 若您想快速体验大模型精调流程，则您无需任何操作可直接使用平台内置示例数据一键启动精调。您可单击此页面的**示例数据**按钮下载查看示例数据内容。
- 若您需要用自定义业务数据精调大模型，则可参考本文档的 [二、挂载 CFS](#) 将业务数据上传到您的 CFS 中。



2. 训练参数配置

单击小卡片详情页的**新建训练任务**按钮后，平台自动为您跳转到**训练工坊 > 任务式建模**的“新建任务”参数配置页面：

2.1 基本信息

←

新建任务

① 状态为运行中的训练任务正在产生费用，不使用时，请及时停止。

基本信息

任务名称

0/256

请输入不超过256个字符，仅支持中英文、数字、下划线“_”、短横“-”，只能以中英文、数字开头

地域

上海

训练镜像

内置大模型 / Llama2 系列模型 / llama_2_7b_chat

训练模式

DDP

机器来源

从CVM机器中选择

从第三方服务器中选择

从T1ONE平台购买

TI-ONE 仅收取软件费用，算力将使用您已购买的机器。[产品定价](#)

资源组

请选择资源组

↻

资源申请

卡型号

请选择卡型号

单节点GPU

—

0.0

+

卡

若需使用GPU，根据不同卡类型可填写0.1-1或1的整数倍。运行环境为平台内置GPU镜像时，卡数不能为0

单节点CPU

—

1.0

+

核

单节点内存

—

1.00

+

G

节点数

—

1

+

个

内置大模型训练资源配置指南请查看 [文档](#)

本页面重点参数配置如下：

- 任务名称：输入您的自定义任务名称。
- 训练镜像：此时平台已自动为您选中“训练框架”为“内置大模型/llama系列/llama_2_7b_chat_hf”。
- 机器来源：选择从 CVM 机器中选择，本示例是采用 A800 的机型配置进行精调。若您还没有资源组，可参考 [资源组管理](#) 产品文档新建。
 - 资源组：选择您采购了 A800 卡机型所在的资源组。
 - 资源申请：CPU 建议40核、内存建议 800G、卡型号建议 A800、GPU建议8卡、节点数建议1个。

2.2 任务配置

存储路径设置 ①

请确保您选择的存储实例（CFS、EMR(HDFS)或者GooseFSx）和纳管资源组的节点网络互通，其中GooseFSx仅支持挂载一个实例

存储类型	用途	CFS文件系统	源路径 ①	容器内路径 ①	操作
CFS	平台内置代码	平台CFS	/code	/opt/ml/code	删除
CFS	平台内置模型	平台CFS		/opt/ml/pretrain_model	删除
CFS	训练输出	请选择	/	/opt/ml/output/data	删除

存储类型	用途	实例	GooseFSx目录 ①	容器内路径 ①	操作
GooseFSx	用户自有数据	请选择	/	/opt/ml/input/data/train	删除

[查看训练数据格式](#)

+ 添加

本页面重点参数配置如下：

- 存储路径设置：默认选中“CFS”（目前大模型精调的平台内置代码、平台内置数据和平台内置模型仅支持 CFS ）。
- 选择存储路径
 - 第一行“平台CFS”：系统默认为您配置了精调该大模型的配套训练代码。
 - 第二行“平台CFS”：
 - 系统默认为您配置了一份精调该大模型的示例数据；

版权所有：腾讯云计算（北京）有限责任公司

第41 共58页

- 若您希望使用自定义业务数据精调该大模型，可删除此行，并在底部添加其他存储来源 CFS 或者 GooseFSx，例如替换为您在第一步中新建的文件系统（如本示例第一步中新建的“llm-sft-demo”），并选择您在第二步中上传的业务数据所在的“源路径”即可。
- 第三行 “平台CFS”：系统默认为您配置了平台内置模型。
- 第四行 “用户CFS”：此处需选择您的 CFS 文件系统（如本示例第一步中新建的“llm-sft-demo”）；“源路径”需填写第二步中新建的子目录如“sft-demo-20230825”；“容器挂载路径”为系统默认填充您无需修改。
- 代码包：无需选择（大模型精调场景中您无需选择代码包）。
- 启动命令：系统默认填充您无需修改（用于启动大模型默认配套的训练代码）。
- 训练输出：无需选择（大模型精调场景中您无需选择训练输出）。
- 调优参数：为系统默认填充，您可按需修改。
 - Epoch: 训练过程中的迭代轮次。
 - BatchSize: 每轮训练迭代中的样本数。BatchSize越大，训练速度越快同时内存占用越高。
 - LearningRate: 梯度下降过程中更新权重的超参，过高导致模型难以收敛，过低导致模型收敛速度过慢。
 - Step: 每跑多少个Step保存一次模型的checkpoint，保存checkpoint越多需要的存储空间越大。
 - UseTlearn: 是否要开启腾讯自研加速，"true/false"。
 - FinetuningType: 用户可自定义选择精调训练模式“Lora/Full”。
 - MaxSequenceLength: 最大文本序列长度，默认“2048”，超过该长度的训练文本将被自动截断。
- 配置完成后，您需单击同意开源大模型精调相关法律协议。最后单击确定按钮即可启动训练。

五、监控训练过程

当任务处于“运行中”状态时，您可单击任务名称，查看任务详情。

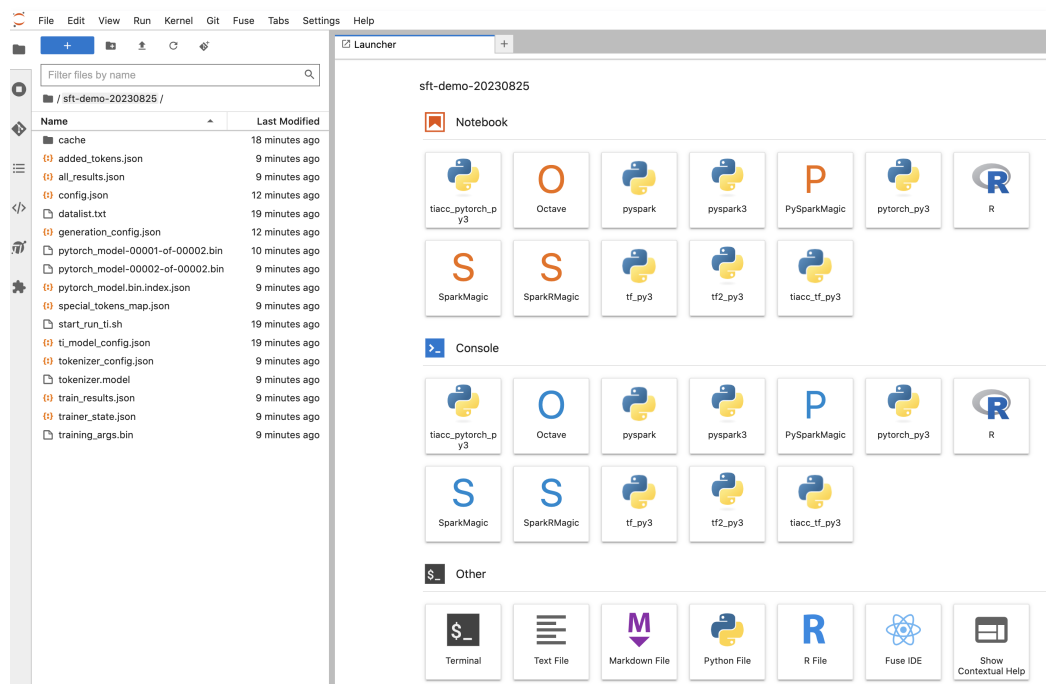
您可切换“基本信息/实例列表/监控/日志/事件”页面查看本次训练任务的详细信息。如监控页面的“资源监控”页面可实时观察资源使用率。

六、查看训练输出文件

若您使用平台内置的示例数据进行精调体验，按本示例的资源配置和参数配置，约 10 分钟即可完成一次精调训练。

当任务状态为“已完成”时，您的业务数据精调大模型的训练任务即成功完成。

精调后得到的大模型存储在您启动训练任务时配置的自定义 CFS 输出路径下。您可通过再次打开第二步中新建的开发机实例，查看 CFS 文件系统上本次训练任务的输出文件列表。



七、部署模型服务

训练成功后，若您想部署精调后的大模型，则可切换到模型服务 > 在线服务二级菜单页面。

在服务列表页面单击新建服务按钮，进入服务参数配置页面：

← 创建服务

服务名称

请输入服务名称

请输入不超过60个字符，仅支持中英文、数字、下划线“_”、短横线“-”，只能以中英文、数字开头

服务版本

V1

服务描述（选填）

请输入不超过500个字

0 / 500

地域

上海

部署方式

☒ 标准部署

☐ 多机分布式部署

☐ 多角色部署(PD分离)

服务新建后，更新服务、新增版本均无法修改部署方式

机器来源

从CVM机器中选择

从第三方服务器中选择

从T1ONE平台购买

TI-ONE 收取软件和机器费用。详见[产品定价](#)

计费模式

☒ 按量计费

☐ 包年包月

副本设置

请确保您选择的存储实例（CFS）和纳管资源组的节点网络互通

模型来源

☐ 模型仓库

☒ CFS

☐ 镜像

选择模型

CFS控制台

运行环境

内置 / LLM(1.0.0, 2.0, 2.4) / angel-vllm(2.1)

开启gRPC

☒

端口

8501

算力规格

8C40G V100*1

大模型推理所需资源指南

高级设置

请求限流

☒ 不限流

☐ 单副本QPS

☐ 单副本最大并发数

副本调节

☒ 手动调节

☐ 自动调节

副本数量

–

1

+

个

配置价格

服务配置费用11.31元/小时

本页面重点参数配置如下：（其余参数都可使用默认值，不修改或按需修改）

- 服务名称：输入您的自定义服务名称。
- 机器来源和计费模式：本示例采用从T1ONE平台购买和按量计费。
- 模型来源：选择“CFS”类型（因为您在配置训练任务时，设置了将精调后的模型存储到自定义 CFS 文件系统中）。
- 选择模型
 - 参数1: 选择存储精调后的大模型的您的 CFS 文件系统名称。
 - 参数2: 输入具体的模型存储路径（此路径和您在训练参数配置时的“源路径”保持一致即可）。
- 运行环境：选择“内置angel-vllm镜像”。
- 算力规格：本示例选择了“8C40G V100*1”，此外还支持A100、A10 机型。
- 最后，同意大模型相关法律协议后，即可单击一键启动服务按钮。

八、服务调用

返回模型服务 > 在线服务列表页面，待大模型服务状态为“运行中”中后，表示服务已成功启动运行。

- 服务调用的两种方式参考第三步中的2.服务调用即可。
- 若您后续使用平台内置示例数据进行精调训练，则可尝试以下问题体验精调前基底模型效果。

测试 prompt1 内容:

```
"Explain the text in one sentence. \n Text:(CNN) -- Actress Lindsay Lohan's father jumped from a third-floor balcony Thursday to try to escape re-arrest, police said, but officers soon caught up with him and took him into custody -- again.\n\nOfficers detained Michael Lohan after his girlfriend, Kate Major, said he violated terms of his release from jail earlier this week by calling her, the Tampa
```


Police Department said in a statement. Investigators arrested Lohan on suspicion of domestic violence after an incident on Monday night, but they let him out of jail after he posted bail.\n\nJust after 1 a.m. Thursday, Lohan's girlfriend called police to say Lohan had violated the terms of his release by calling her, the police statement said. Lohan called his girlfriend again while she was talking with police, and she put him on speakerphone, it said.\n\nAfter prosecutors authorized Lohan's re-arrest, officers went to a hotel, the Tahitian Inn, where he was staying.\n\n"Upon seeing officers, he jumped out of a third-story balcony in an attempt to escape arrest," the police statement said.\n\nOfficers took him into custody after a short chase and discovered after taking him to jail that he may have broken his foot, police said. Doctors were evaluating him.\n\nLohan is not expected to be released from Tampa General Hospital Thursday and has been admitted for overnight observation, said Tampa police spokeswoman Andrea Davis.\n\nPolice responded to a domestic violence call at Michael Lohan's home Monday and arrested him for the battery of his live-in girlfriend, police said.\n\nLohan told reporters Wednesday he "didn't lay a hand" on his girlfriend.\n\n"I did not hurt her," he said.\n\nLohan also said he never was served with a restraining order to stay away from the woman. According to CNN Tampa affiliate WFTS, a Sarasota County judge issued a temporary restraining order Tuesday.\n\nLohan, 51, was arrested in Los Angeles in March in connection with alleged domestic violence and was charged with one misdemeanor count of corporal injury to a cohabitant.\n\nThe relationship between Lohan and his daughter has been publicly strained for years, although the two did undergo family counseling together during her treatment at the Betty Ford Center.\n\nCNN's Rich Phillips contributed to this report.\n\n"

测试 prompt2 内容:

"Explain the text in one sentence. \n Text:(CNN) -- A federal judge Friday ruled in favor of a former UCLA college basketball star who sued to end the NCAA's control over the rights to college athletes' names, images and likenesses.\n\nIn a landmark decision, U.S. District Judge Claudia Wilken sided with Ed O'Bannon in his lawsuit against the National Collegiate Athletic Association. O'Bannon argued athletes in the top tier of college basketball and football should be allowed to profit from their schools' use of their likenesses.\n\nIn a 99-page ruling, Wilken wrote that current NCAA rules "unreasonably restrain trade in the market for certain educational and athletic opportunities offered by NCAA Division I schools." \n\nWilken issued an injunction to block the NCAA from prohibiting its member schools and conferences from offering their Football Bowl Subdivision or Division I basketball recruits a limited share of the revenues generated from the use of their names, images, and likenesses.\n\nShe did rule, however, that the NCAA could set a cap on the money paid to athletes, as long as it allows at least \$5,000 per athlete per year.\n\n"The NCAA's witnesses stated that their concerns about student-athlete compensation would be minimized or negated if compensation was capped at a few thousand dollars per year," the judge wrote.\n\nO'Bannon's suit alleged the waivers the athletes are required to sign are illegal and asked that players be able to collectively negotiate the terms of their likenesses in order to keep a share of those profits.\n\n"Before the court in this case is only whether the NCAA violates antitrust law by agreeing with its member schools to restrain their ability to compensate Division I men's basketball and FBS football players any more than the current association rules allow," Wilken wrote. \n\nFor the reasons set forth above, the court finds that this restraint does violate antitrust law.\n\nThe ruling could potentially change college sports drastically, eventually forcing the NCAA to restructure its amateur model and allow college athletes to be paid.\n\nNCAA chief legal officer Donald Remy said: "We disagree with the court's decision that NCAA rules violate antitrust laws. We note that the court's decision sets limits on compensation, but are reviewing the full decision and will provide further comment later." \n\nWilliam Isaacson, an attorney for the plaintiffs, called the ruling "a big step forward for common decency." \n\nOne of the things the judge is saying here .. is some sharing is OK," he said. \n\nIt won't affect amateurism, won't affect the popularity of the sport. She made a very reasonable and significant and measured decision." \n\nRamogi Huma, president of the National College Players Association, said the ruling was "a big win" that signaled "the time for college athletes to get their due," though he was critical of the \$5,000 compensation cap.\n\nThe ruling says the NCAA was operating illegally and college athletes do have rights," said Huma, who helped find players to join the case. \n\nEven if you label them student-athletes and want to call it amateurism, it doesn't give the NCAA the right to deny them the rights that other Americans deserve." \n\nSonny Vaccaro, who started the lawsuit by introducing O'Bannon and lead attorney Michael Hausfeld, said the decision was precedent-setting.\n\nThe key is, they're allowed to get paid," said Vaccaro, who helped pioneer branding athletes by putting Nike shoes on Michael Jordan.\n\nVaccaro called the decision his most important career accomplishment.\n\nIt's more important to me that these kids won and this go forward and the principles are right," he said. \n\nTo me, it's more important than the Jordan and Kobe and things I did in my professional life. ...This


```
was just something that was wrong and I totally believed in it."
The judge wrote that the injunction will not affect student-athletes who enroll in college before July 1, 2016.
When athletes commit to a university, players are required to sign a waiver that relinquishes their right to their own likenesses in every form.
That means they can't make money off their television appearances, their jerseys, or in any other way.
The universities get any revenues from selling sports paraphernalia or other material related to the players.
The trial began June 9 in federal court in Oakland, California. The plaintiffs were 20 current and former student athletes who play or played for an FBS football or Division I men's basketball team starting in 1956.
Legal appeals could delay a final outcome for years but the decision is in a position to be the first major NCAA reform effort to take hold.
Already the issues brought up in the case have had an effect, even before the ruling was made.
Texas A&M, the University of Arizona and Northwestern University have decided to stop selling jerseys with the numbers of specific players.
Instead, Texas A&M will sell the number 12 jersey, in keeping with its 12th man tradition; and Arizona will sell jerseys with numbers that correspond to the year of competition -- 14 for this year, according to a school spokesman. Northwestern will sell only jersey number 51, in honor of its head coach, Pat Fitzgerald, and legendary Chicago linebacker Dick Butkus.
The NCAA's argument in both the ongoing O'Bannon suit and another one filed by former quarterback Sam Keller, also in federal court in Oakland, is that it is trying to protect the amateur model of college sports.
Paying college athletes would hurt traditions, NCAA chief testifies
NCAA under fire: 5 things to know
NCAA faces change, legal challenges in months ahead
```

一些精调经验的建议

如何选择基底模型

建议您可以参考各类模型的技术报告，了解模型在开源评测集中的指标，从参数量以及重要评测指标两个角度来选择最适合自己的模型。

首先我们需要决定使用多大参数量的模型。目前，主流的自然语言大模型都是 decode-only 的结构，训练显存占用与训练速度都能够比较容易地推算（基本上和参数量成正比）。7B模型的最低资源要求是1台8卡32G显存V100，13B模型的最低资源要求是1台8卡40G显存A100，70B模型的最低资源要求是4台8卡40G显存A100。训练完成后，使用fp16推理部署，7B模型需要1卡32G显存V100，13B模型需要1卡40G显存A100，70B模型需要4卡40G显存A100。用户可以根据自己的训练与推理资源，选择参数量。原则上，参数量越大的模型，效果会有较大的提升。

然后我们可以根据模型的开源评测指标或者对自己的业务测试集进行自测摸底，来确定使用具体的哪个模型。一般来说，国外的开源模型例如llama3通用能力包括语言理解、指令遵循、数学计算等整体都比较出色，但对中文的支持会有所欠缺。而国内的开源模型，由于进行了中文文化的训练，对中文的支持都比较好。

关于使用 Base 模型还是 Chat 模型作为基座模型，这两个选择有各自的优点：

1. 使用Base模型能够保证自己的数据不会和开源的指令精调数据发生冲突，例如重复、开源模型的某项能力不佳等。通过这种方式用户能够更好地把控模型的效果。
2. 使用 Chat 模型用户可以迁移开源模型所训好的能力，当用户数据不够充足的情况下，可以使用 Chat 模型更快速地达到效果。

我们提供了国内外多款主流的开源模型，包括 llama3, baichuan, qwen 等，同时也提供了我们腾讯云的行业大模型。我们腾讯云行业大模型通过大量行业语料增训，使得模型具备强大的行业领域知识，同时也具备较强的阅读理解能力，70B的模型效果能够与市面上主流的闭源模型效果相当。

如何选择精调方式

TI-ONE 提供全量参数精调与 LoRA 的方式。这两种方式的主要区别为：

1. 全量参数精调的上限比较高，能够取得更好的业务指标，但是对训练数据质量、数量的要求以及计算资源的需求较高。如果训练数据质量、数量不够，全量参数精调会大幅破坏模型的原有基础任务能力，只倾向于拟合训练数据。
2. LoRA 能够较好保持原有模型的能力，需要的计算资源也较少（根据 LoRA 的 rank 超参，训练时间大约为全量参数精调的1/2-1/10）。但是模型能力的上限较低，受限于训练时选用的基座模型。

因此，当数据量比较少（低于1万），且任务类型比较单一，推荐使用 LoRA 的方式进行精调。对于在大模型领域有经验的同学，同时拥有丰富的质量训练数据，建议使用全量参数精调的方式。

如何配置超参

用户启动训练时，需要关心序列长度、批量大小（batch size）、学习率、迭代轮数（epoch 数）

- 序列长度：在训练代码中，会根据用户设置的序列长度，对训练数据进行截断。因此用户需要根据自己的训练数据的最长序列长度进行设置，以保证某些重要的数据不会被错误截断。为了提高训练效率，用户可以将不同长度的数据进行分批次训练（例如先训4k再训8k）。在设置序列长度时，也建议用户对基座模型进行核对，看是否支持相应的序列长度。
- 批量大小：一般来说，用户在计算资源（主要是显存）允许的条件下，尽可能提升批量大小。该批量大小和模型参数大小以及序列长度高度相关，也将很大程度上影响训练时间。此外，也建议在训练中要保证有足够的迭代次数（1000步以上），迭代次数由数据量、迭代轮数、批量大小来决定。在数据量较小，又不适合过多地增加迭代轮数，则可以适当减少批量大小。

- 学习率：初始学习率建议选择 $5e-6$ 至 $1e-5$ 之间。如果多次开展增量训练，学习率需要相应地减小。如果loss下降很慢，可以适当的增大，如果过拟合比较严重，可以适当降低学习率。
- 迭代轮数：建议1-5个 epoch，需要兼顾总迭代次数（1000步以上）。因此如果数据量较小（例如几千条），可以适当增大迭代轮数（不建议）。推荐增加数据的多样性，提升数据量，而不是简单的多训几个 epoch。建议数据量达到1万-10万的量级，迭代轮数为1-2个 epoch 为佳。

增量训练技巧：

由于数据存在序列长度，以及难易程度的区别，可以考虑增量训练的方式，先基于基座模型训练模型A，然后在A基础上继续训练出模型B。

在增量训练的过程中需要注意以下两点：

- 因为模型训练会存在遗忘的问题，因此在第二阶段训练中，需要搭配一部分与第一阶段训练同分布的数据。具体方式可以是第一阶段训练数据中采样部分，也可以是在第一阶段训练中有分离出来一部分同分布数据没有加入训练。这样模型在第二阶段训练的学习过程中，不仅能学习新数据，同时也能维持住第一阶段模型A的能力。
- 第二阶段的初始学习率要比第一阶段的学习率要低，例如第一阶段为 $1e-5$ ，第二阶段可以选用 $5e-6$ 。

多阶段增量训练的这种方式可以用于先训练短文本、再训练长文本；先训练较难的任务，例如数学计算，再训练全量的数据。实验证明较难的任务建议多训练几个 epoch，可以通过这种两阶段的方式达到目的。

传统 AI 模型系列

使用任务式建模构建手写体分类模型

最近更新时间：2024-08-22 17:42:31

任务式建模简介

任务式建模提供通过向导式的训练任务提交方式进行模型构建，支持基于多种算法来源进行训练任务提交，可直接通过代码包绑定主流训练框架启动训练任务，快速使用主流高性能及分布式训练框架提交训练任务。下面将由一个简单的 PyTorch MPIJob 演示如何使用任务式建模快速创建任务。

数据准备

数据集

本案例使用 mnist 数据集，下载地址为 [数据集](#)。

代码包

本案例的训练脚本是使用 PyTorch 框架撰写的，代码包下载地址为 [代码包](#)。

上传数据到 COS 对象存储中

您可以进入 [COS 控制台](#)，在存储桶列表页面创建存储桶，详情请参考 [创建存储桶](#)。

创建好的存储桶将用于平台任务数据的存放，包括数据集、代码包等，文件上传操作详情请参考 [上传对象](#)。

⚠ 注意

1. COS 为对象存储产品，独立计费，详细可见 [对象存储-计费概述](#)。
2. 创建 COS 存储桶时，所属地域需跟训练任务所在的地域一致，如训练任务的地域处于广州，创建存储桶时所属地域应选择广州。

操作步骤

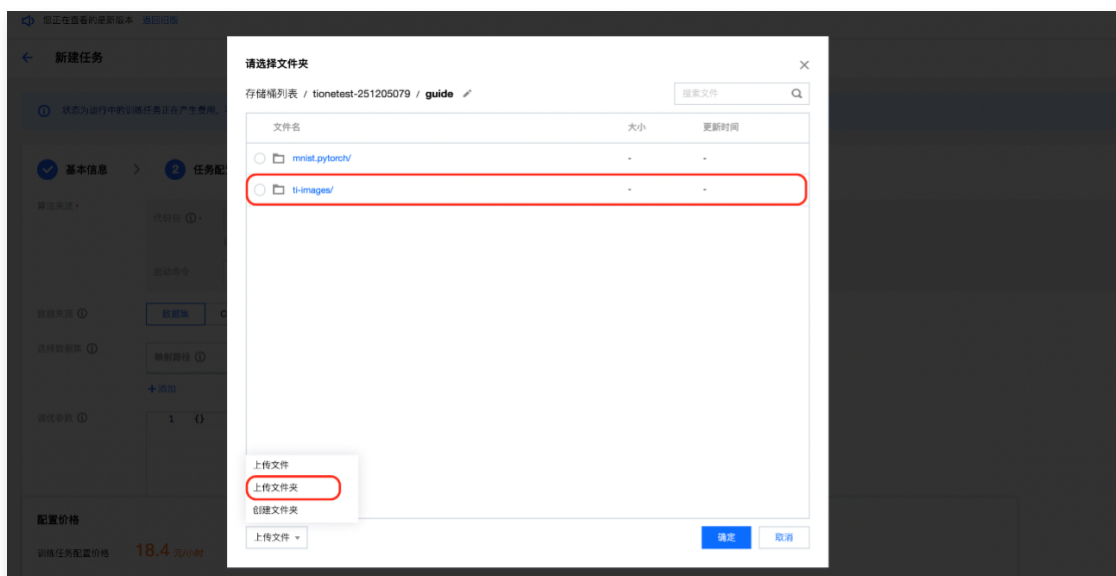
新建任务第一步

1. 进入训练工坊 > 任务式建模，单击新建，开始进入向导式训练任务创建。
2. 在基本信息页，填写如下信息：
 - 任务名称：mnist_train
 - 训练镜像选择：内置镜像 / PyTorch / torch1.9-py3.8-cuda11.1-gpu
 - 训练模式：MPI
 - 计费模式：按量计费
 - 算力规格：8C40G V100*1
 - 节点数量：1个
 - 标签和描述：无需填写

新建任务第二步

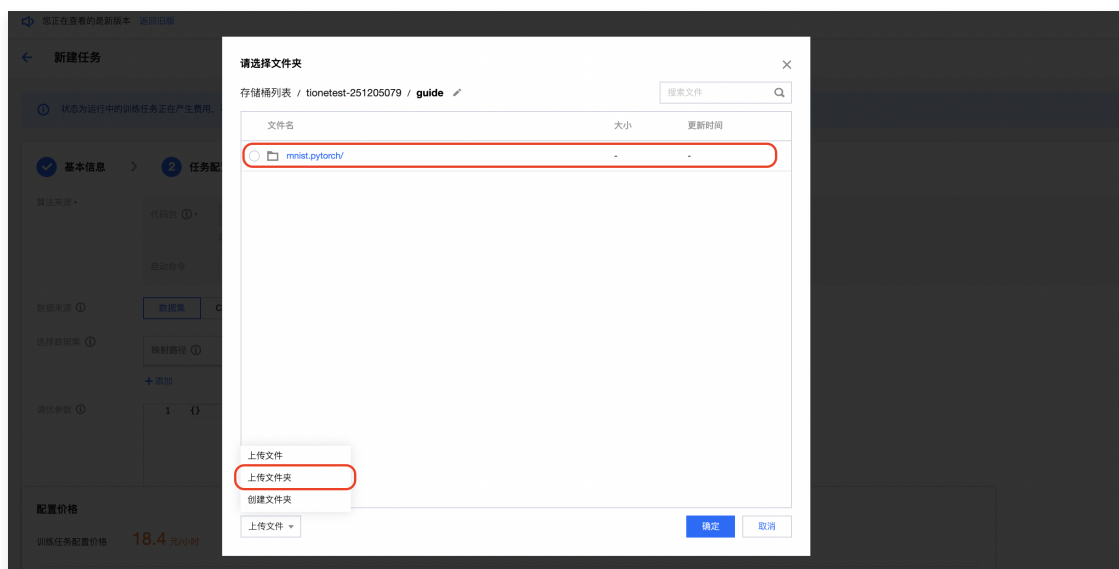
在任务配置页，填写如下信息：

1. 数据配置：选择 COS 数据
 - 本地存储路径：填写 train
 - 数据所在路径：单击选择文件，在弹出的COS对话框中，选择需要使用的存储桶，单击左下方上传文件夹，将数据集解压后的文件夹ti-images上传，上传完成后选中文件夹路径，如下图所示：



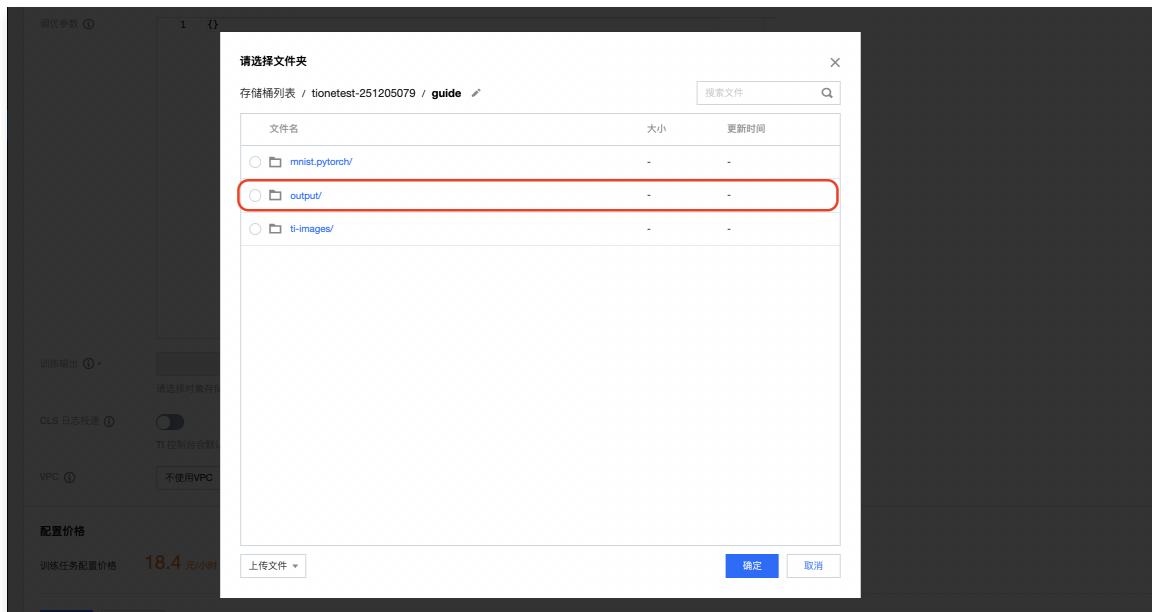
2. 代码包:

单击**选择文件**，在弹出的 COS 对话框中，选择需要使用的存储桶，单击左下方**上传文件夹**将准备好的代码包（需要先解压）文件夹mnist.pytorch上传至COS存储桶中，并选定代码包所在路径。



3. 启动命令: 填写 sh start.sh

4. 训练输出：单击**选择文件**，在弹出的COS对话框中，选择需要使用的存储桶，选择训练输出数据需要保存的路径，如下图所示：



5. 调优参数：无

6. 私有化网络：无

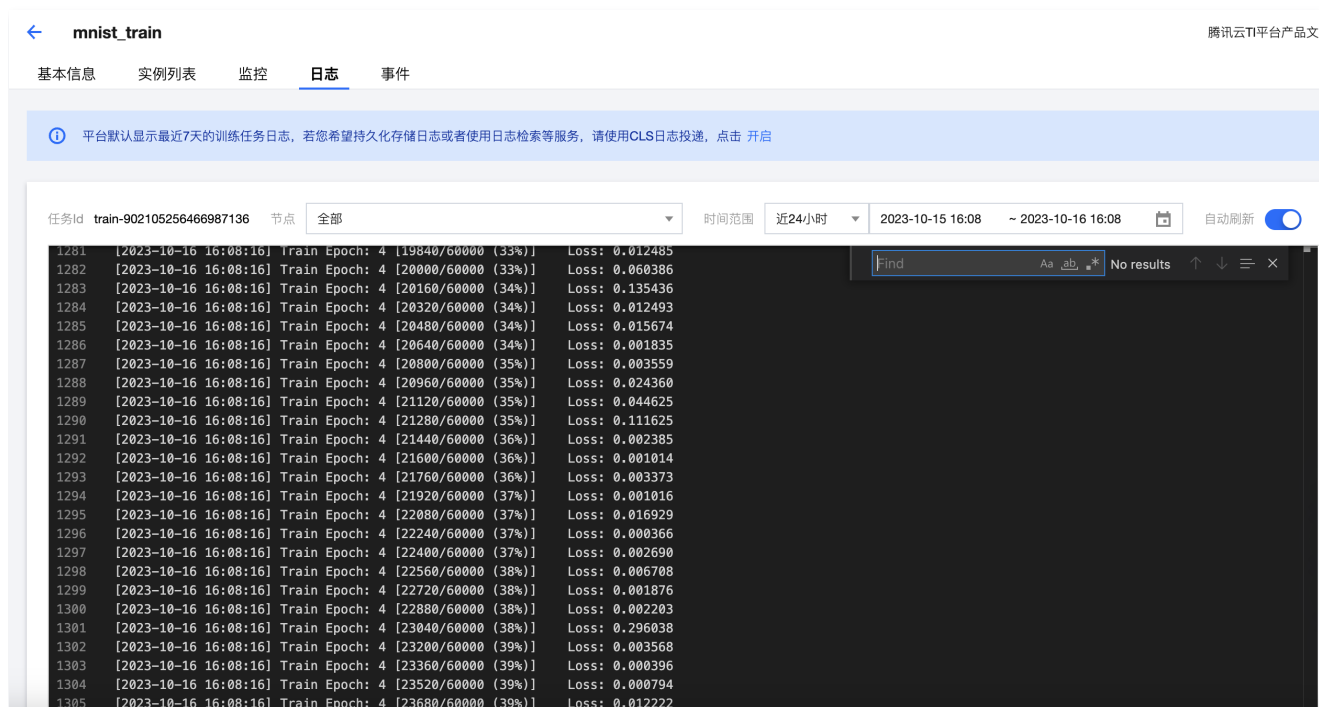
7. CLS 日志：选择不投递

配置完成后，可在页面底部查看本次训练任务的每小时收费价格，单击**确定**，即完成任务提交。

查看和监控任务

1. 提交成功后，可在任务列表页面看到任务记录。

2. 点击任务名称，可进入任务详情页查看日志和监控信息，例如下图所示：



导入和部署第三方 AI 模型

最近更新时间：2024-05-14 17:42:31

本文档包含1个案例，案例描述了将第三方模型部署为在线服务的使用链路，适用于仅需使用 TI-ONE 进行模型推理的用户。

案例：将第三方模型导入平台，并部署为在线服务

本案例以平台预置模型包中 PyTorch 图像分类模型为例，讲解如何将第三方模型部署为在线服务，部署完成后，用户可通过在线测试功能调用服务，识别输入图片的图像种类。

准备内容

1. 模型包

1.1 登录 [TI-ONE 控制台](#)，在左侧导航栏中选择**模型管理** > **模型仓库**，进入模型列表页面；

1.2 下载对应的模型文件和推理文件 demo，可 [单击此处](#) 下载。

1.3 [这里](#)以pytorch > detect模型为例演示后续步骤。

2. 测试数据已为您准备了如下测试数据，可以在服务发布后直接用于调用测试, 可 [单击此处](#) 下载。

步骤1：导入模型仓库

1. 进入 [TI-ONE 控制台](#)，在左侧导航栏中选择 **模型管理** > **模型仓库**，进入模型列表页面。

2. 单击 **导入模型**，在模型导入页面参考如下参数填写。

- 导入方式：选择 **导入新模型**
- 模型名称：自定义模型名称，例如命名为 detection
- 标签：按需添加，可不配置
- 模型来源：选择**从 COS 导入**（需要提前将模型文件上传到 COS 中）
- 模型格式：选择 **TorchScript**
- 运行环境来源：建议使用内置的运行环境，例如 pytorch1.9.0-py38(cpu)
- 模型文件：单击**选择文件**，在弹出的 COS 对话框中，选择需要使用的存储桶，单击左下方**上传文件夹**，将模型包中的**模型文件夹**上传，上传完成后选中文件夹路径

导入模型

导入方式: **导入新模型** | 导入新版本 | 导入至现有版本

模型名称: ✓
请输入不超过60个字符，仅支持中英文、数字、下划线"_"、短横"-"，只能以中英文、数字开头

标签: + 添加

模型来源: **从任务导入** | **从COS导入**

模型格式: **TorchScript**

运行环境来源: **内置 / pytorch1.9.0-py38(cpu)**

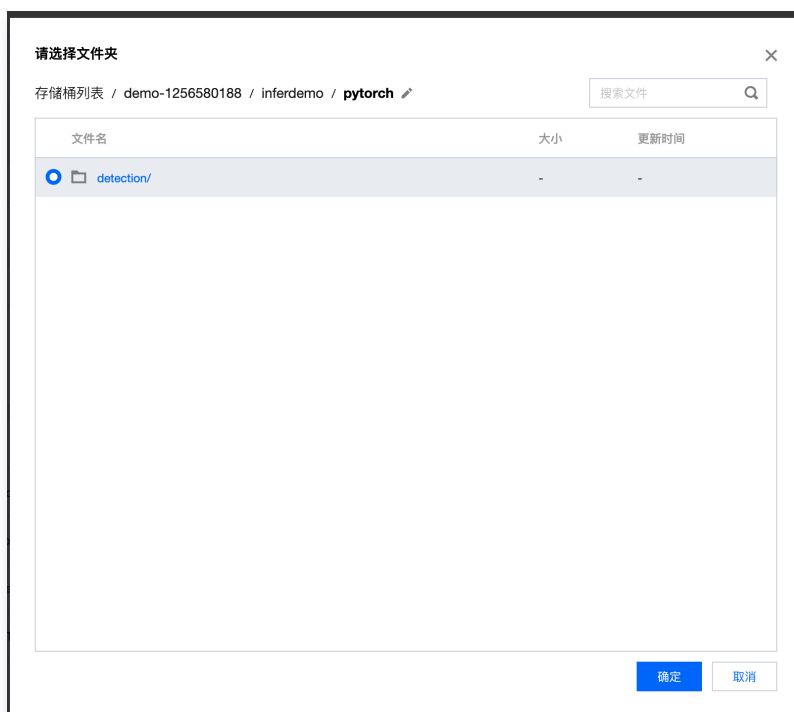
算法框架: **PyTorch**

其他配置

模型指标:

模型来源路径: 选择文件路径

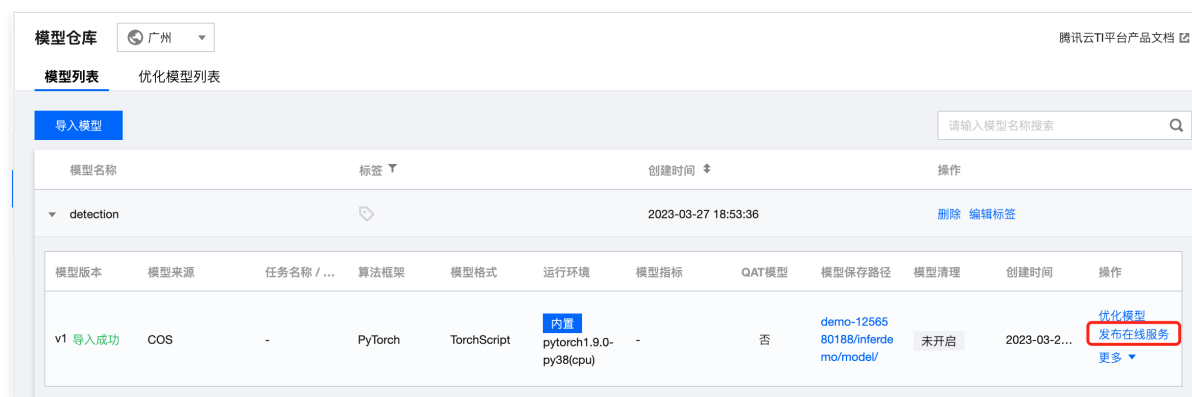
确定 **取消**



3. 单击确定，完成模型文件的导入。

步骤2：启动在线服务

1. 进入 [TI-ONE 控制台](#)，在左侧导航栏中选择 **模型管理 > 模型仓库**，选择步骤1创建好的模型。
2. 单击 **操作-发布在线服务**，在启动服务页面参考如下参数，填写服务基础信息



- 服务名称：自定义服务名称，例如命名为 classify
- 服务描述：按需填写服务描述信息，可不配置
- 计费模式：若您尚未使用过资源组管理模块，建议选择 **按量付费** 模式

3. 服务实例信息，可参考如下参数填写：

- 模型来源：默认选择 **从模型仓库选择模型**
- 选择模型：默认选择上一步骤中导入的模型文件 detection
- 选择版本：默认选择 v1
- 模型推理文件：将模型推理文件 model_service.py 上传（下载模型包中已经提供了 model_service.py，会自动带出，用户也可以自行上传）
- 存储挂载：默认不填写
- 算力规格：按需选择，例如 32C128G T4 * 1
- 高级配置/启动命令：默认不填写

○ 高级配置/环境变量：默认不填写

服务实例

模型来源

☒ 模型仓库
 ☐ CFS
 ☐ 镜像

此选项将导入您在模型仓库中注册的模型，您可以前往[模型仓库](#)注册模型，目前模型仓库仅支持TorchScript、Detectron2、ONNX、Frozen Graph、Saved Model、MMDetection、Hugging Face-bert、Hugging Face-stable diffusion、Hugging Face-stable diffusion(动态加速)、PMMLL格式的模型

模型类型

通用

加速

选择模型

detection

选择版本

v1

运行环境

tione.tencentcloudcr.com/qcloud-ti-platform/ti-cloud-infer-pytorch:py38-tc

运行环境

环境版本

py38-torch1.9.0-cu111-tiacc4.1.1-3.2.0

选择版本

模型推理文件

model_service.py

重新上传

提示：模型推理文件定义了推理逻辑，需要按照平台的规范开发 [示例下载](#)。建议发布前先在本地进行模型推理文件开发和调试，步骤如下
 1. 安装开发环境：python init_infer_env.py --framework torch1.9.0 --model-format TorchScript --model-scene detect --demo-dir ./（建议您在独立的python虚拟环境中安装）。点击下载 [init_infer_env.py文件](#)
 2. 开发调试：本地开发调试详细指南请查看 [文档](#)
[收起](#)

存储挂载

CFS文件系统

CFS控制台

CFS源路径

算力规格

32C128G T4*1

高级设置

4. 参考如下信息完成服务高级配置

- 实例调节：按需选择，可使用默认值 **手动调节**
- 实例数量：按需配置，可使用默认值1
- 是否生成鉴权：按需开启，可使用默认值关闭
- CLS 日志投递：按需开启，可使用默认值关闭
- 请求限流：按需开启，可使用默认值关闭
- 是否开启 TI-ONE 内网访问外部：按需开启，可使用默认值关闭
- 自动停止：按需开启，可使用默认值关闭
- 标签：按需添加，可不配置

实例调节

☒ 手动调节
 ☐ 自动调节

实例数量

-

1

+

个

是否生成鉴权

☐ 开启鉴权后，服务调用时会进行签名认证，已启动的服务可在服务调用页面查看签名密钥及签名计算指引

CLS 日志投递

☐ TI 控制台会默认展示 7 天的日志，若您期望持久化存储日志，获得日志检索等服务，可以开启 CLS 日志投递，CLS 产品介绍和收费指南请查看[文档](#)

请求限流

☐

是否开启TI-ONE内网访问外部

☐ 开启后配置VPC和子网，会在您账户下的孩子网，创建一个弹性网卡进行绑定，独立计费，弹性网卡当前未计费

自动停止

☐ 开启后，在线服务将在指定的停止时间自动停止，同时停止服务算力计费

标签

[+ 添加](#)

5. 单击启动服务，若选择按量计费模式，会出现费用冻结确认。按量计费模式需要预先冻结两小时费用，若您账户内余额充足，单击确定即可完成服务创建。此时在服务列表中，新创建的服务会处于**创建中**状态，服务部署过程中将为您创建网关并调度计算资源，需要等待一段时间，待服务成功完成部署时，服务状态将变为**运行中**。

步骤3：服务调用测试

1. 确认上一步中发布的服务已完成部署，处于 **运行中** 的状态。

2. 单击 调用 操作，进入服务调用页面。

在线服务

广州

腾讯云TI平台产品文档

新建服务

请输入名称搜索

名称	状态	计费模式	服务ID	运行中/总版本数	标签	创建时间	操作
detection 计费中	运行中	按量付费	ms-9pggp7x	1/1	-	2023-03-27 19:24:46	新增版本 编辑标签 调用 删除

3. 在接口信息页面，可直接进行在线测试，本模型的示例请求如下：

- 在请求体中录入图片路径

```
{ "image": "https://tione-prod-open-1256580188.cos.ap-guangzhou.myqcloud.com/ti-cloud/ti-ems/tiacc-demo-models/ts-yolov5/dining_table.jpg" }
```

此时模型会实时返回测试结果。

接口信息

接口调用地址

http://service-a7x9ln7q-1256580188.gz.apigw.tencentcs.com:80/tione/v1/models/m:predict

服务类型

HTTP

请求方法

POST

调用方式(命令行)

curl -X POST http://service-a7x9ln7q-1256580188.gz.apigw.tencentcs.com:80/tione/v1/models/m:predict -H 'Content-Type:application/json' -d ''

调用方式(在线测试)

请求体(Request Body 600KB 内)

```
1 {
2   "image": "https://tione-prod-open-1256580188.cos.ap-guangzhou.myqcloud.com/ti-cloud/ti-ems/tiacc-demo-models/ts-yolov5/dining_table.jpg"
3 }
```

发送请求

请求响应(Response)

```
1 Status: 200 OK
2 Connection: keep-alive
3 Content-Length: 1321
4 Date: Mon, 27 Mar 2023 11:33:03 GMT
5 Server: openresty/1.21.4.1
6 X-API-Id: api-bg6gxsfe
7 X-API-Requestid: 1b80e11fa043ee567d1e8a5ede0447c5
8
9 {
10   "result": {
11     "det_objs": {
12       "det_boxes": [
13         [
14           289,
15           297,
16           481,
17           461
18         ],
19         [
20           293,
21           294,
22           481,
```

Tikit 快速入门

最近更新时间：2025-05-14 17:03:52

TiKit 简介

TiKit 是 TI-ONE 提供的开源 Python SDK 软件包。用户可以通过 TiKit 使用 TI-ONE 的各项功能，例如提交和查看训练任务。

安装

腾讯云 TI 平台环境内（包括 Notebook 和训练任务的容器）已经安装了 TiKit，可以直接开始使用。

非腾讯云 TI 平台环境内，安装方法如下：

安装依赖

目前 TiKit 可以运行在 3.6 及以上的版本的 Python3 环境中。

您可以按照如下步骤快速创建一个能够使用 TiKit 的容器（需自行准备 docker 环境），或者参考如下步骤准备好符合要求的环境。

ubuntu:20.04

1. 启动 ubuntu:20.04 容器：

```
docker run -ti --network=host ubuntu:20.04 /bin/bash
```

2. 安装依赖包：

```
sudo apt-get update
sudo apt-get install -y libsasl2-dev libkrb5-dev python3-dev python3-pip
```

安装 TiKit

- 方式一：使用 pip 安装（推荐）

```
pip3 install -U tikit
```

- 方式二：离线安装。在 <https://pypi.org/project/tikit/> 上下载安装包，使用 whl 文件安装，或者使用源码安装：

```
pip3 install tikit-1.0.0-py3-none-any.whl
```

```
# 或者，解压源码后
python3 setup.py install
```

查看版本

```
pip3 show tikit | grep Version
```

开始使用

- 登录腾讯云，打开 [API 密钥管理页面](#)，点击“新建密钥”按钮，保存弹窗展示的 SecretId，SecretKey。
- 打开 Python3 环境，使用上一步获取的密钥创建客户端：

腾讯云 TI 平台环境

```
import tikit
client = tikit.client.Client("${SecretId}", "${SecretKey}")
```

非腾讯云 TI 平台环境

```
import tikit
client = tikit.Client("${SecretId}", "${SecretKey}", "${Region}")
```

目前支持的地区如下：

地区	Region
广州	ap-guangzhou
上海	ap-shanghai
上海自动驾驶云	ap-shanghai-adc
南京	ap-nanjing
北京	ap-beijing

基本操作

查看命令帮助

以查看提交训练任务的命令 `create_training_task` 的帮助为例：

```
help(client.create_training_task)
```

查看内置训练框架

通过 `describe_training_frameworks` 可以查看平台内置训练镜像和框架：

Notebook

在 Notebook 中，运行命令即可自动打印结果。

```
client.describe_training_frameworks()
```

```
[2]: client.describe_training_frameworks()
```

[2]:	框架名称	版本	训练模式
	PYTORCH	tilearn-llm1.0-torch2.1-angel-vllm1.0-py3.10-cuda12.1-gpu	DDP
	PYTORCH	tilearn-llm0.4.2-torch2.1-deepspeed0.10.0-py3.10-cuda12.1-gpu	DDP
	PYTORCH	ti-acc2.0-torch1.9-py3.8-cuda11.1-gpu	DDP
	PYTORCH	torch1.9-py3.8-cuda11.1-gpu	DDP, MPI, HOROVOD
	PYTORCH	torch1.9-deepspeed0.8.2-py3.8-cuda11.1-gpu	DDP, MPI, HOROVOD
	PYTORCH	torch2.1-py3.8-npu	DDP, MPI, HOROVOD
	TENSORFLOW	ti-acc1.0-tf1.15-py3.6-cuda10.0-gpu	PS_WORKER
	TENSORFLOW	tf1.15-py3.7-cpu	PS_WORKER, MPI, HOROVOD
	TENSORFLOW	tf1.15-py3.7-cuda10.0-gpu	PS_WORKER, MPI, HOROVOD
	TENSORFLOW	tf2.4-py3.8-cpu	PS_WORKER, MPI, HOROVOD
	TENSORFLOW	tf2.4-py3.8-cuda11.1-gpu	PS_WORKER, MPI, HOROVOD
	SPARK	spark2.4.5-cpu	SPARK
	PYSPARK	spark2.4.5-py3.7-cpu	SPARK

终端

在普通终端中，可以通过 `print` 函数打印命令的结果。

```
print(client.describe_training_frameworks())
```

```
>>> print(client.describe_training_frameworks())
```

框架名称	版本	训练模式
PYTORCH	tilearn-llm1.0-torch2.1-angel-vllm1.0-py3.10-cuda12.1-gpu	DDP
PYTORCH	tilearn-llm0.4.2-torch2.1-deepspeed0.10.0-py3.10-cuda12.1-gpu	DDP
PYTORCH	ti-acc2.0-torch1.9-py3.8-cuda11.1-gpu	DDP
PYTORCH	torch1.9-py3.8-cuda11.1-gpu	DDP, MPI, HOROVOD
PYTORCH	torch1.9-deepspeed0.8.2-py3.8-cuda11.1-gpu	DDP, MPI, HOROVOD
PYTORCH	torch2.1-py3.8-npu	DDP, MPI, HOROVOD
TENSORFLOW	ti-acc1.0-tf1.15-py3.6-cuda10.0-gpu	PS_WORKER
TENSORFLOW	tf1.15-py3.7-cpu	PS_WORKER, MPI, HOROVOD
TENSORFLOW	tf1.15-py3.7-cuda10.0-gpu	PS_WORKER, MPI, HOROVOD
TENSORFLOW	tf2.4-py3.8-cpu	PS_WORKER, MPI, HOROVOD
TENSORFLOW	tf2.4-py3.8-cuda11.1-gpu	PS_WORKER, MPI, HOROVOD
SPARK	spark2.4.5-cpu	SPARK
PYSPARK	spark2.4.5-py3.7-cpu	SPARK

提交训练任务

1. 创建训练镜像配置。以提交使用内置镜像 `tilearn-llm0.4.2-torch2.1-deepspeed0.10.0-py3.10-cuda12.1-gpu` 为例，该镜像的框架名称为 `PYTORCH`，训练模式为 `DDP`：

```
framework = tikit.models.FrameworkInfo.new_system_framework("PYTORCH", "tilearn-llm0.4.2-torch2.1-deepspeed0.10.0-py3.10-cuda12.1-gpu", "DDP")
```

2. 创建共享存储配置。以挂载类型为 CFS 的共享存储的训练任务为例（备注说明：以下以 CFS 为例进行详细说明，平台也支持挂载 GooseFSx 和 GooseFS）：

- `${FileSystemID}`：CFS 文件系统 ID。请打开 [腾讯云 CFS 文件系统控制台](#) 获取同地域可用的 CFS 文件系统 ID，或者向您的运维人员咨询可用 CFS 文件系统 ID。并通过 Notebook 或者参考 CFS 操作文档，将您的代码、数据、模型等物料放入该 CFS 文件系统的某个目录中。
- `${SourcePath}`：挂载源路径。请填写上述 CFS 文件系统中存放了您的代码、数据、模型等物料的目录的绝对路径。
- `${TargetPath}`：挂载目标路径。填写您希望在训练容器中看到上述物料的目录的绝对路径。

```
data_config = tikit.models.TrainingDataConfig.new_mount_cfs("${FileSystemID}", "${SourcePath}",
"${TargetPath}")
```

3. 创建资源配置。

后付费

从 TIONE 平台购买，并且按量计费。通过 `describe_postpaid_training_price` 可以查看可以购买的机型以及价格。
以申请 2 核 CPU、2GB 内存 的 `TI.S6.MEDIUM2.POST` 机型为例：

```
resource = tikit.models.ResourceConfigInfo.new_postpaid("TI.S6.MEDIUM2.POST")
```

预付费

使用预付费资源组，包括从 CVM 选择，以及从 TIONE 平台购买的包年包月的资源组。请打开 [资源组管理页面](#) 获取同地域可用的资源组 ID，或者向您的运维人员咨询可用资源组 ID。

以使用资源组 `${ResourceGroupID}`，1 核 CPU、1GB 内存、1 卡 `A100` 为例：

```
resource = tikit.models.ResourceConfigInfo.new_prepaid(cpu=1, memory=1, gpu=1, gpu_type="A100")
resource_group_id = "${ResourceGroupID}"
```

4. 设置训练任务名称和命令：

- `${TaskName}`：自定义任务名称。请注意任务不可重名。
- `${StartCmd}`：任务运行的 shell 命令。例如休眠 10s 后退出 `sleep 10`。

```
task_name = "${TaskName}"
start_cmd="${StartCmd}"
```

5. 创建训练任务。

`cos_output` 和 `code_package_path` 为当您希望将代码和训练输出存储在 [腾讯云对象存储](#) 时，使用该参数，当前例子下，置空即可。

```
client.create_training_task(task_name, \
    framework=framework, \
    cos_output="", \
    worker_resource=resource, \
    code_package_path="", \
    input_data_config=[data_config], \
    worker_start_cmd=start_cmd, \
    resource_group_id=resource_group_id)
```

命令执行成功后，将打印训练任务的 ID。输出示例：

```
{"Id": "train-1234564123451234520", "RequestId": "xxx"}
```

