

MoPaaS 大模型平台操作手册

上海殷泊信息科技有限公司

2024 年 09 月

第 1 章 引言

本文以 ChatGLM2、Stable Diffusion 和 Llama2 的示例进行阐述，分别从数据准备→大模型精调实验→管理大模型→大模型服务的流程进行介绍。

第 2 章 ChatGLM2

2.1 准备数据

1. 执行以下操作前需要准备文件 chatglm2.zip, 如下图:



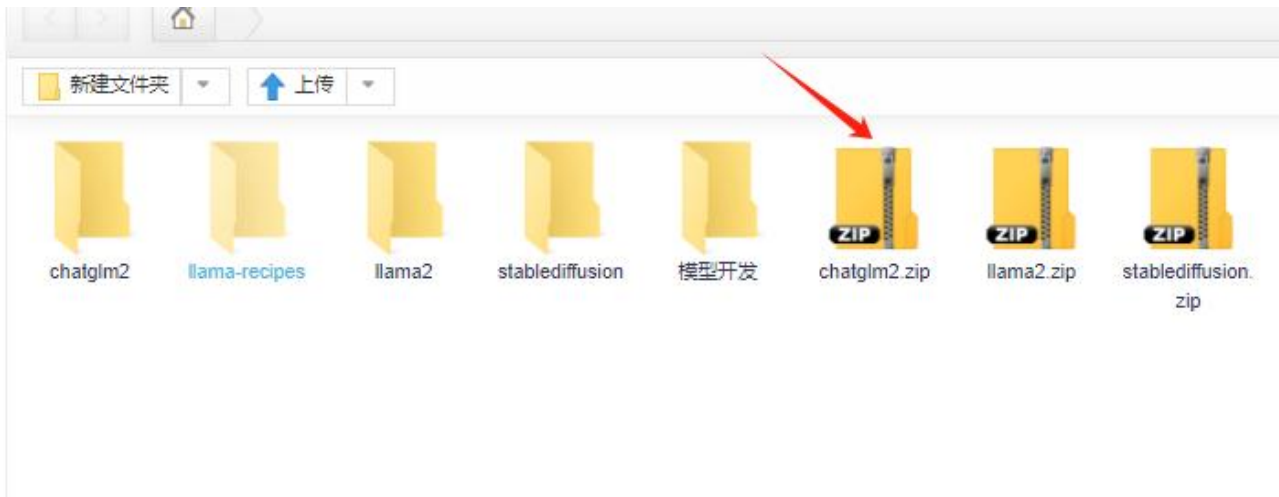
2. 新建存储

登录后进入【我的资产→文件管理】，新建存储，待存储状态为正常。

☐	名称/ID	状态	已使用/总量	是否共享	创建时间	用户	操作
☐	data aa1e0fa08cdc	● 正常	907.09 MB / 2.00 GB	44.29% 否	2023-02-07 10:03:00	demo01 demo01	详情 管理 共享 更多

3. 上传训练数据和代码

点击【管理】打开存储管理页面，上传文件 chatglm2.zip 并解压。



2.2 大模型精调实验

1. 创建实验

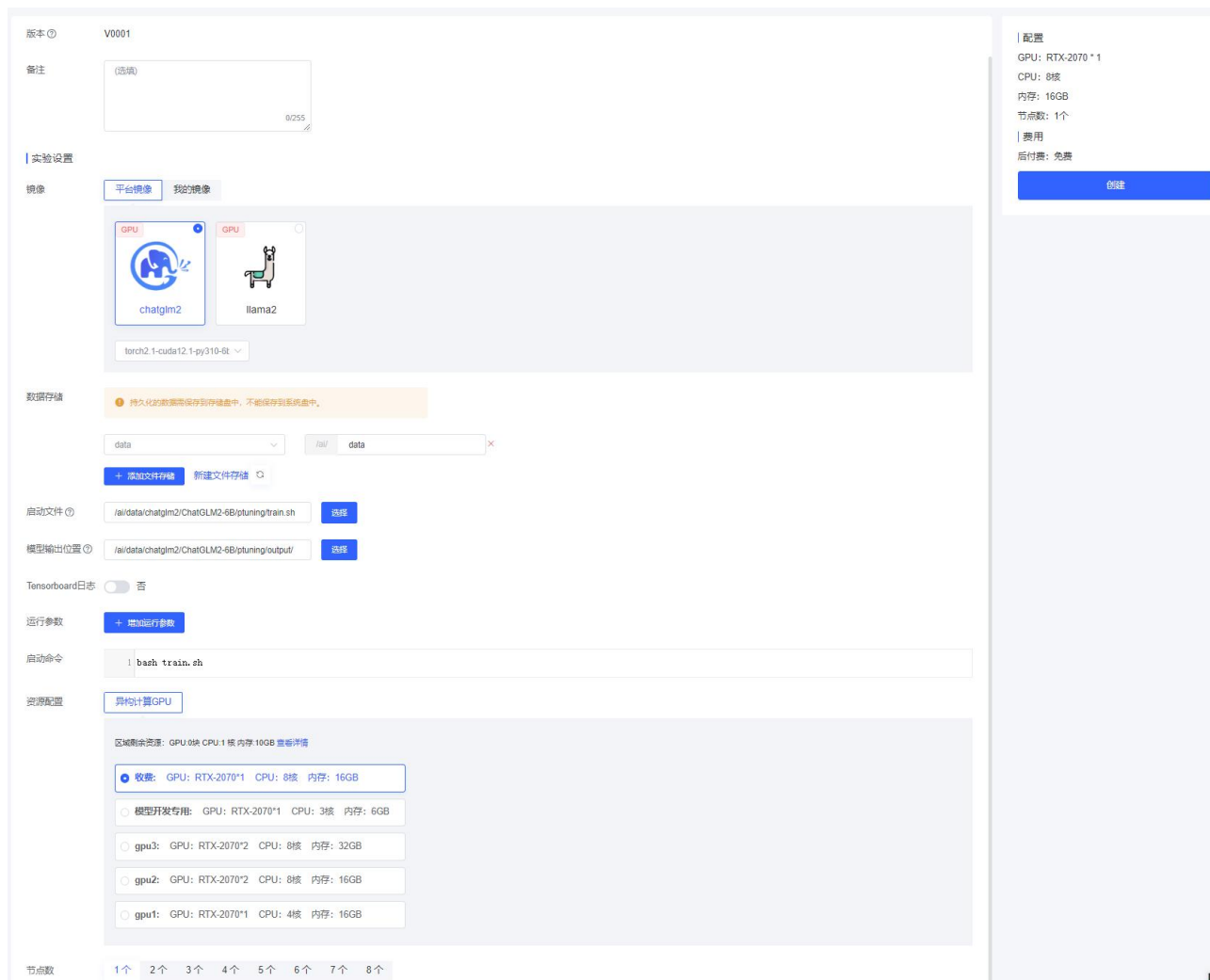
进入【模型调优→LLM 实验】，点击【创建实验】跳转至创建实验页面。

填写基础信息：名称（符合格式要求即可）。

进行实验设置：镜像选择 chatglm2、版本选择 torch2.1-cuda12.1-py310-6b、选择文件存储并输入挂载目录(data)、选择启动文件 (/ai/data/chatglm2/ChatGLM2-6B/ptuning/train.sh)、选择模型输出位置 (/ai/data/chatglm2/ChatGLM2-6B/ptuning/output/)、资源配置选择异构计算 GPU 任意一个单卡套餐、节点数选择 1 个，其他均保持默认选项。

设置完成后，点击【创建】即可。

实验设置参考下图所示：



2. 开始训练

实验创建成功后，若资源充足，会自动进行训练，状态变为【训练中】。

点击【实验名称】或者【详情】，进入实验详情页后可查看训练日志。

版本信息							
版本	状态	镜像	配置	训练时长	优先级	创建时间	操作
V0003 9784b3853eed	训练中	chatglm2:torch2.1-cuda12.1-py310-6b	gpu1-4-14 节点数:1 GPU:V100*1 CPU:4核 内存:14GB	12分15秒	--	2024-09-04 10:56:04	详情 停止 生成子版本 日志 更多

2.3 管理大模型

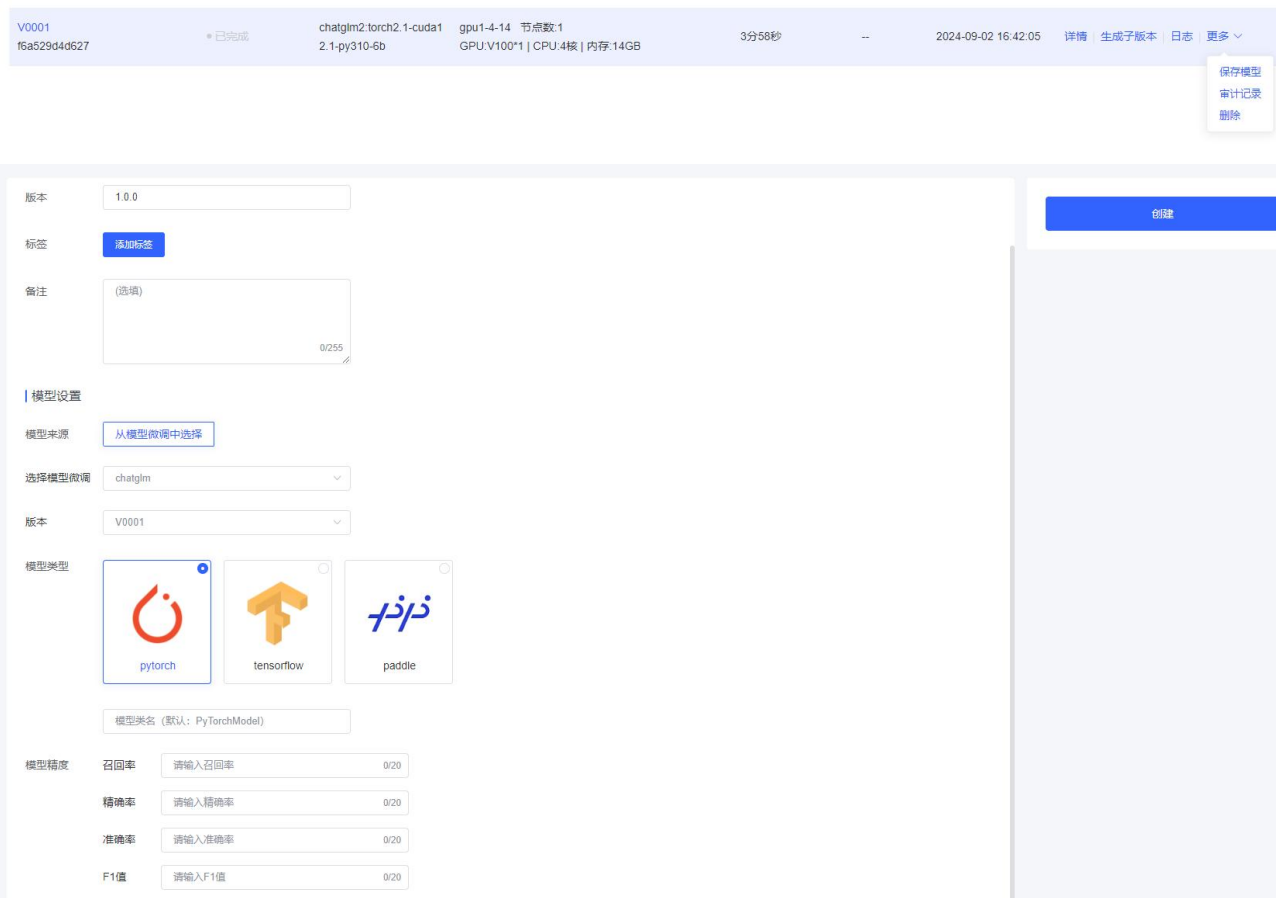
待实验训练完成，在实验详情页面，点击【更多】→【保存模型】进入保存模型页面。

填写基础信息：名称（符合格式要求即可）。

进行模型设置：选择模型微调和版本，类型选择 pytorch，其他均保持默认选项。

设置完成后，点击【创建】，待模型状态为导入成功，则模型已成功导入。

导入模型设置参考下图所示：



导入成功如下图所示：

☐	名称/ID	版本	状态	模型类型	模型来源	创建时间	用户	操作
☐	chatglm2 2aab3535eb8c	1.0.0	● 导入成功	pytorch	模型微调: chatglm/V0001	2024-09-04 11:12:24	用户头像	详情 新增版本 发布为服务 更多

2.4 大模型服务

1. 发布服务

进入【模型服务→在线服务】，点击【发布服务】跳转至发布服务页面。

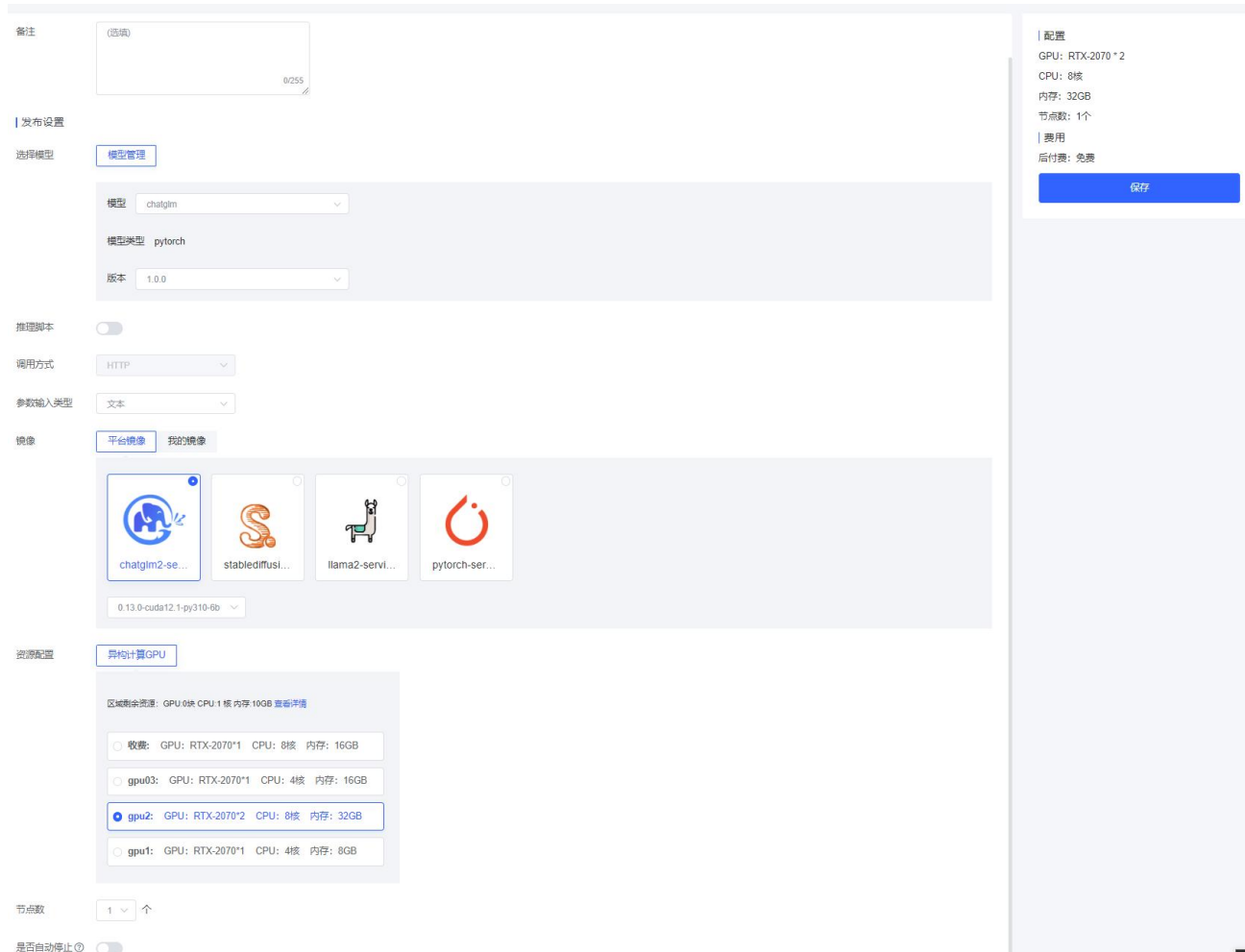
填写基础信息：名称（符合格式要求即可）。

进行发布设置：选择上一章成功导入的模型和版本，镜像选择 chatglm2-serving、版本选择

0.13.0-cuda12.1-py310-6b，资源配置选择任意一个套餐、节点数选择 1 个，其他均保持默认选项。

设置完成后，点击【创建】即可。

发布服务参考下图所示：



备注: (可选) 0/255

发布设置

选择模型

模型: chatglm

模型类型: pytorch

版本: 1.0.0

推理脚本: ☐

调用方式: HTTP

参数输入类型: 文本

镜像

平台镜像 我的镜像

chatglm2-se... stablediffusi... llama2-servi... pytorch-ser...

0.13.0-cuda12.1-py310-6b

资源配置

异构计算GPU

区域剩余资源: GPU: 0块 CPU: 1核 内存: 10GB [查看详情](#)

☐ 收藏: GPU: RTX-2070*1 CPU: 8核 内存: 16GB

☐ gpu03: GPU: RTX-2070*1 CPU: 4核 内存: 16GB

☒ gpu2: GPU: RTX-2070*2 CPU: 8核 内存: 32GB

☐ gpu1: GPU: RTX-2070*1 CPU: 4核 内存: 8GB

节点数: 1 个

是否自动停止: ☐

配置

GPU: RTX-2070 * 2

CPU: 8核

内存: 32GB

节点数: 1个

费用

后付费: 免费

保存

2. 对话

在线服务发布成功后待其为【运行中】状态，在推理服务列表点击【对话】。

☐	名称ID	状态	配置	调用成功次数/总次数	运行时长	创建时间	用户名/姓名	操作
☐	chatglm 514ac38dc765	运行中	gpu2 节点数: 1 GPU: RTX-2070*2 CPU: 8核 内存: 32GB	3/4	2天15小时7分24秒	2024-08-27 16:57:06	III III	详情 停止 对话 更多

对话

×



输入你的问题，有问必答

写一段歌词



Verse 1:阳光洒在沙滩上 让我感到温暖 音乐在心中回荡 让我感到快乐

Chorus:让我们荡起双桨 向着未来前进 让我们勇敢地拼搏 走向明天

Verse 2:朋友在身边陪伴 让我感到快乐 梦想在心中燃烧 充满希望与梦想 让我们勇往直前

Chorus:让我们荡起双桨 向着未来前进 让我们勇敢地拼搏 走向明天

Bridge:手握着水中的花 感受着它的美丽 脚下的海浪 声声地和我对话

Chorus:让我们荡起双桨 向着未来前进 让我们勇敢地拼搏 走向明天

Outro:让我们荡起双桨 向着未来前进 让我们勇敢地 走向明天

输入你想了解的内容

提交

第3章 Stable Diffusion

3.1 准备数据

1. 执行以下操作前需要准备文件 stablediffusion.zip, 如下图:



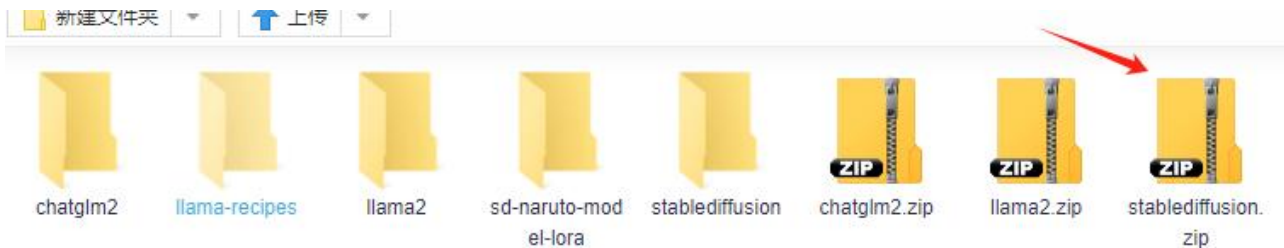
2. 新建存储

登录后进入【我的资产→文件管理】，新建存储，待存储状态为正常。

☐	名称/ID	状态	已使用/总量	是否共享	创建时间	用户	操作
☐	data aa1e0fa08cdc	正常	907.09 MB / 2.00 GB	44.29% 否	2023-02-07 10:03:00	demo01 demo01	详情 管理 共享 更多

3. 上传训练数据和代码

点击【管理】打开存储管理页面，上传文件 stablediffusion.zip 并解压。



3.2 大模型精调实验

1. 创建实验

进入【模型调优→文生图实验】，点击【创建实验】跳转至创建实验页面。

填写基础信息：名称（符合格式要求即可）。

进行实验设置：镜像选择 stablediffusion、版本选择 v1.4-torch2.3-cuda12.1-py310、选择文件存储并输入挂载目录(data)、选择启动文件 (/ai/data/stablediffusion/train.sh)、选择模型输出位置 (/ai/data/stablediffusion/sd-naruto-model-lora/)、资源配置选择异构计算 GPU 任意一个套餐、节点数选择 1 个，其他均保持默认选项。

设置完成后，点击【创建】即可。

实验设置参考下图所示：

版本 ① V0001
备注 (可选)

0/255

实验设置

平台镜像

我的镜像

GPU

stablediffusion

v1.4-torch2.3-cuda12.1-py310

持久化的数据需保存到存储盘中，不能保存到系统盘中。

data

/ai/ data

+ 添加文件存储

新建文件存储

启动文件 ①

/ai/data/stablediffusion/train.sh

选择

模型输出位置 ①

/ai/data/stablediffusion/sd-naruto-model-lora/

选择

Tensorboard日志

否

运行参数

+ 添加运行参数

启动命令

1 bash train.sh

资源配置

异构计算GPU

区域剩余资源: GPU:3块 CPU:9核 内存:45GB [查看详情](#)

gpu2-6-30: GPU: V100*2 CPU: 6核 内存: 30GB

gpu1-4-14: GPU: V100*1 CPU: 4核 内存: 14GB

节点数

1个 2个 3个 4个 5个 6个 7个 8个

配置

GPU: V100 * 1

CPU: 4核

内存: 14GB

节点数: 1个

费用

后付费: 免费

创建

2. 开始训练

实验创建成功后，若资源充足，会自动进行训练，状态变为【训练中】。

点击【实验名称】或者【详情】，进入实验详情页后可查看训练日志。

版本信息							
版本	状态	镜像	配置	训练时长	优先级	创建时间	操作
V0004 767b80f60331	训练中	stablediffusion v1.4-torch h2.3-cuda12.1-py310	gpu1-4-14 节点数:1 GPU:V100*1 CPU:4核 内存:14GB	31秒	--	2024-09-09 14:50:11	详情 停止 生成子版本 日志 更多

3.3 管理大模型

待实验训练完成，在实验详情页面，点击【更多】→【保存模型】进入保存模型页面。

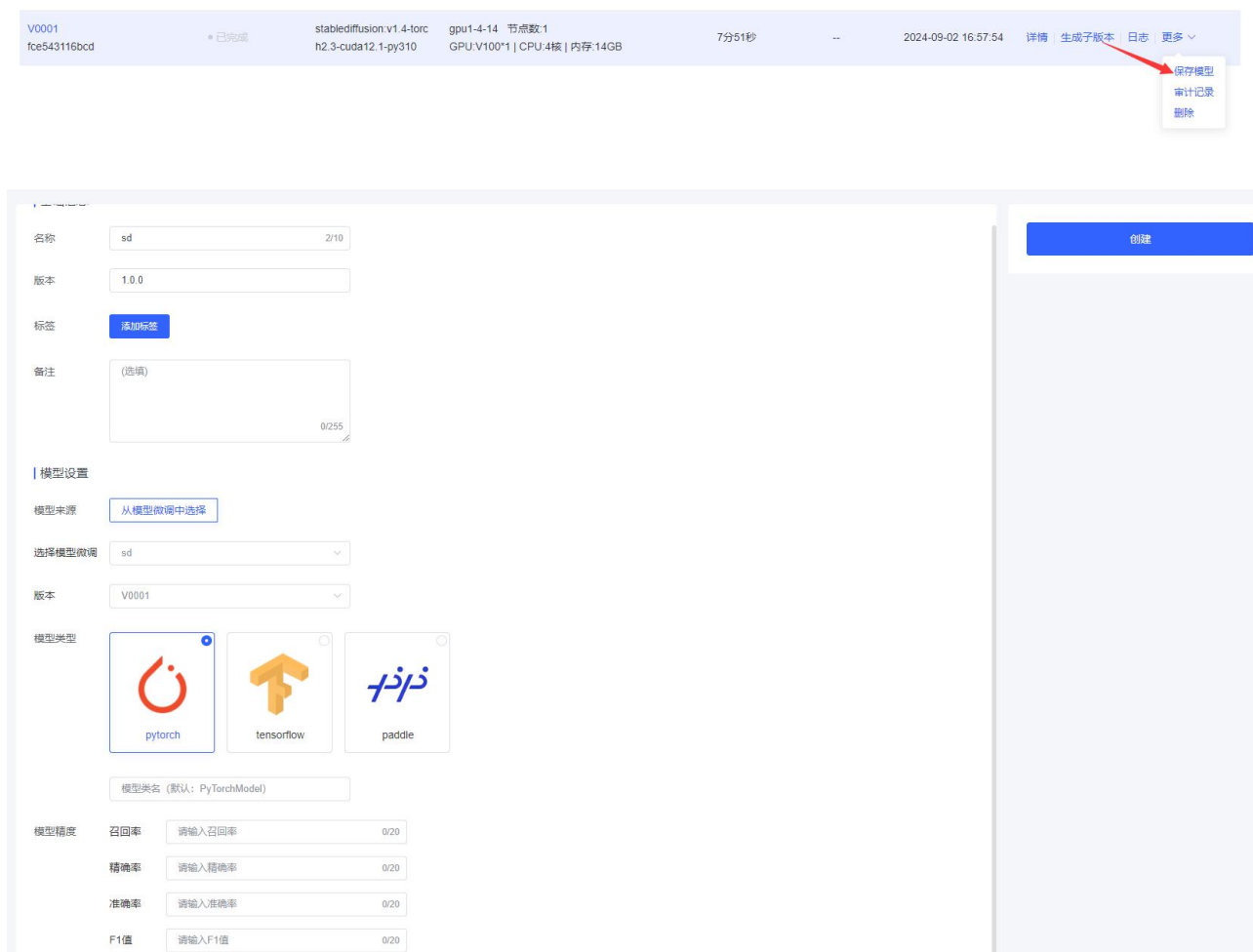
填写基础信息：名称（符合格式要求即可）。

进行模型设置：选择模型微调和版本，类型选择 pytorch，其他均保持默认选项。

设置完成后，点击【创建】，待模型状态为导入成功，则模型已成功导入。

导入模型设置参考下图所示：

8



导入成功如下图所示:



3.4 大模型服务

1. 发布服务

在线服务列表，点击【发布服务】跳转至发布服务页面。

填写基础信息：名称（符合格式要求即可）。

进行发布设置：选择上一章成功导入的模型和版本，镜像选择 `stablediffusion-serving`、版本选择 `v1.4-torch2.3-cuda12.1-py310`，资源配置选择任意一个套餐、节点数选择 1 个，其他均保持默认选项。

设置完成后，点击【创建】即可。

发布服务参考下图所示:

备注

(选项)
0/255

| 发布设置

选择模型

模型管理

模型

sd2

模型类型

pytorch

版本

1.0.0

推理脚本

调用方式

HTTP

参数输入类型

文本

镜像

平台镜像

我的镜像

AI

chatglm2-se...

stablediffusi...

llama2-servi...

pytorch-ser...

v1.4-torch2.3-cuda12.1-py31i

资源配置

异构计算GPU

区域剩余资源: GPU:3块 CPU:9核 内存:45GB [查看详情](#)

单卡: GPU: V100*1 CPU: 4核 内存: 15GB

gpu02: GPU: V100*2 CPU: 8核 内存: 30GB

gpu: GPU: V100*2 CPU: 8核 内存: 32GB

节点数

1 个

是否自动停止

1个小时

2个小时

3个小时

6个小时

12个小时

24个小时

配置

GPU: V100 * 1

CPU: 4核

内存: 15GB

节点数: 1个

费用

后付费: 免费

创建

2. 对话

在线服务发布成功后待其为【运行中】状态，在推理服务列表点击【对话】。

☐	sd2 49df394d6510	运行中	单卡 节点数:1 GPU:V100*1 CPU:4核 内存:15GB	0/0	19小时15分47秒	2024-09-02 17:22:15		详情 停止 对话 更多
--------------------------	---------------------	-----	---	-----	------------	---------------------	--	---

10

对话

×

AI

请输入你的提示语，一键为你创作图像

(chubby panda: 1.2) ,bamboo forest, cute, mascot, <lora:blindbox_V1Mix:1>, (8k, RAW photo, best quality),(masterpiece),(best quality),(ultra-detailed), (full body:1.2), masterpiece:1.2,((Flower background))

人

AI



第 4 章 Llama2

4.1 准备数据

1. 执行以下操作前需要准备文件 llama2.zip, 如下图:



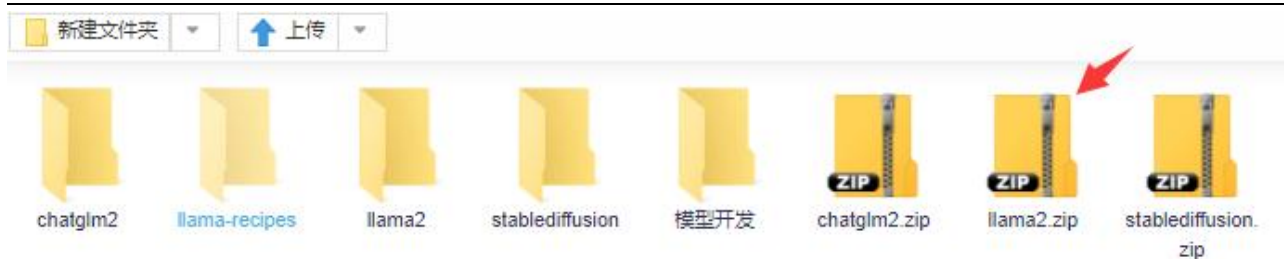
2. 新建存储

登录后进入【我的资产→文件管理】，新建存储，待存储状态为正常。

☐	名称/ID	状态	已使用/总量	是否共享	创建时间	用户	操作
☐	data aa1e0fa08cdc	• 正常	907.09 MB / 2.00 GB	44.29% 否	2023-02-07 10:03:00	demo01 demo01	详情 管理 共享 更多

3. 上传训练数据和代码

点击【管理】打开存储管理页面，上传文件 llama2.zip 并解压。



4.2 大模型精调实验

1. 创建实验

进入【模型调优→LLM 实验】，点击【创建实验】跳转至创建实验页面。

填写基础信息：名称（符合格式要求即可）。

进行实验设置：镜像选择 llama2、版本选择 v5、选择文件存储并输入挂载目录(data)、选择启动文件（ /ai/data/llama2/llama-recipes/llama_finetuning.py ） 、 选 择 模 型 输 出 位 置（/ai/data/llama-recipes/output/）、添加运行参数、资源配置选择异构计算 GPU 任意一个套餐、节点数选择 1 个，其他均保持默认选项。

设置完成后，点击【创建】即可。

实验设置参考下图所示：

0/255

实验设置

镜像

平台镜像 我的镜像

GPU AI chatglm2

GPU llama2

v5

数据存储

持久化数据需保存到存储盘中，不能保存到系统盘中。

data /ai/ data

+ 添加文件存储 新建文件存储

启动文件

/ai/data/llama2/llama-recipes/llama_finetuning.py 选择

模型输出位置

/ai/data/llama-recipes/output/ 选择

Tensorboard日志

否

运行参数

use_peft = 请输入参数值 请输入备注

peft_method = lora 请输入备注

quantization = 请输入参数值 请输入备注

model_name = /root/.cache/huggingface/hub/Llama-2-7b-hf 请输入备注

output_dir = /ai/data/llama-recipes/output/ 请输入备注

dataset = alpaca_dataset 请输入备注

num_epochs = 1 请输入备注

+ 增加运行参数

启动命令

1 python llama_finetuning.py --use_peft --peft_method=lora --quantization --model_name=/root/.cache/huggingface/hub/Llama-2-7b-hf --output_dir=/ai/data/llama-r

资源配置

异构计算GPU

区域剩余资源: GPU:1块 CPU:15核 内存:61GB 查看详情

gpu2-6-30: GPU: V100*2 CPU: 6核 内存: 30GB

gpu1-4-14: GPU: V100*1 CPU: 4核 内存: 14GB

节点数

1个 2个 3个 4个 5个 6个 7个 8个

配置

GPU: V100 * 1

CPU: 4核

内存: 14GB

节点数: 1个

费用

后付费: 免费

创建

运行参数如下：

use_peft

peft_method=lora

Quantization

model_name=/root/.cache/huggingface/hub/Llama-2-7b-hf

output_dir=/ai/data/llama-recipes/output/

dataset=alpaca_dataset

num_epochs=1

2. 开始训练

实验创建成功后，若资源充足，会自动进行训练，状态变为【训练中】。

点击【实验名称】或者【详情】，进入实验详情页后可查看训练日志。

版本	状态	镜像	配置	训练时长	优先级	创建时间	操作
V0003 cd62672a69e2	训练中	llama2-v5	gpu1-4-14 节点数:1 GPU:V100*1 CPU:4核 内存:14GB	41秒	--	2024-09-04 13:53:11	详情 停止 生成子版本 日志 更多

4.3 管理大模型

待实验训练完成，在实验详情页面，点击【更多】→【保存模型】进入保存模型页面。

填写基础信息：名称（符合格式要求即可）。

进行模型设置：选择模型微调和版本，类型选择 pytorch，其他均保持默认选项。

设置完成后，点击【创建】，待模型状态为导入成功，则模型已成功导入。

导入模型设置参考下图所示：

V0001 be0fe97f2373	已完成	llama2-v5	gpu1-4-14 节点数:1 GPU:V100*1 CPU:4核 内存:14GB	14小时21分4秒	--	2024-09-02 17:26:12	详情 生成子版本 日志 更多
-----------------------	-----	-----------	--	-----------	----	---------------------	----------------

保存模型
审计记录
删除

导入成功如下图所示：

名称ID	版本	状态	模型类型	模型来源	创建时间	用户	操作
llama2 e80faeeef178b	1.0.0	导入成功	pytorch	模型源: llama2-v0001	2024-08-27 14:25:15	llm	详情 新增版本 发布为服务 更多

4.4 大模型服务

1. 发布服务

在线服务列表，点击【发布服务】跳转至发布服务页面。

填写基础信息：名称（符合格式要求即可）。

进行发布设置：选择上一章成功导入的模型和版本，镜像选择 llama2-serving、版本选择 0.13.0-cuda12.1-py38，资源配置选择任意一个套餐、节点数选择 1 个，其他均保持默认选项。

设置完成后，点击【创建】即可。

发布服务参考下图所示：

在线服务 / 发布服务

0/255

发布设置

选择模型

模型管理

模型 llama2

模型类型 pytorch

版本 1.0.0

推理脚本 ☐

调用方式 HTTP

参数输入类型 文本

镜像

平台镜像 我的镜像

AI chatglm2-se...

S stablediffusi...

llama2-servi...

pytorch-ser...

0.13.0-cuda12.1-py38

资源配置

异构计算GPU

区域剩余资源: GPU:1块 CPU:15 核 内存:61GB 查看详情

单卡: GPU: V100*1 CPU: 4核 内存: 15GB

gpu02: GPU: V100*2 CPU: 8核 内存: 30GB

gpu: GPU: V100*2 CPU: 8核 内存: 32GB

节点数 1 个

是否自动停止 ☒

1个小时

2个小时

3个小时

6个小时

12个小时

24个小时

配置

GPU: V100 * 1

CPU: 4核

内存: 15GB

节点数: 1个

费用

后付费: 免费

创建

2. 对话

在线服务发布成功后待其为【运行中】状态，在推理服务列表点击【对话】。