1 比赛指导文档

1.1 华为云申请代金券指南

在华为云平台训练需要使用代金券，领取方式见下文。注意代金券数量有限，先到先得，代金券金额有限，**请节约使用，并及时关注余额，避免欠费。**

操作方式：

Step1：代金券申请

首先登陆华为云，链接：https://auth.huaweicloud.com/authui/login.html?locale=zh-cn&service=https%3A%2F%2Fwww.huaweicloud.com%2F#/login，如果已经有华为云账号可直接登陆，如果没有需要先注册账号，然后实名认证。注册完华为云账号之后，需要进行全局配置，操作如下图：

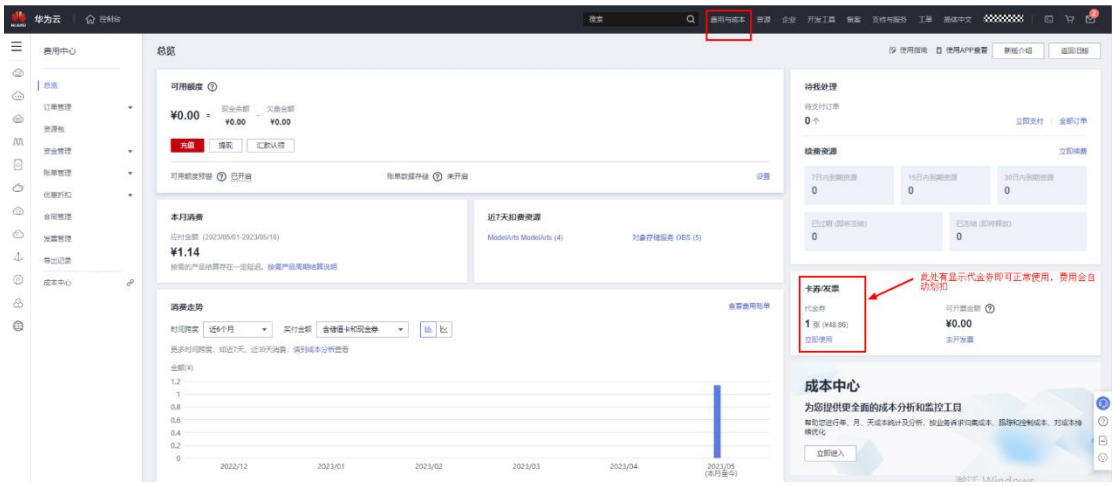
配置完成以后不要做其他操作（额外操作可能会收取费用导致账号欠费，需手动充值），去领取华为云代金券，注意代金券金额有限请谨慎使用。代金券领取链接详见比赛的各赛题官网页面，进入链接以后按照要求填写相关信息，提交申请。

Step2 代金券发放：

审核标准：1. 选手需报名参加对应赛题；2. 申请选手需为队伍队长；

代金券到账时会进行短信提醒，同时可通过此链接查看代金券是否到账：https://account.huaweicloud.com/usercenter/?agencyId=0e15c42d26c14cef994ead1af42648f9&region=cn-north-4&locale=zh-cn#/userindex/allview

打开界面如下图所示：



【特别提醒】

请参赛团队及时关注代金券额度，如发现额度较少，请先停止训练、删除服务。

1、由于比赛会用到昇腾算力、OBS存储等，会产生少量费用，因此在进行比赛操作前务必领取代金券，按照操作手册操作，以免账号欠费。代金券仅能在激活的账户上使用，参赛队员可与各自团队队长详细沟通代金券激活账户信息。

2、领取代金券资源后，请仔细了解代金券涵盖的资源类型，对于不包含的资源类型，或超出资源规格将会产生费用；

3、代金券到期后，如需继续使用相关服务，将产生相应费用。请在比赛结束后，及时删除不需要的项目，防止因资源到期产生不必要的扣费。释放资源请点击链接了解详情：

https://support.huaweicloud.com/usermanual-billing/renewals_topic_70000001.html

4、训练完成后，注意观察ModelArts首页是否还有计费中服务，并及时进行关闭；

5、您创建大赛所需资源时会优先扣除已领取的按需代金券，超出部分以按需付费的方式进行结算。如果您使用了其他类型规格的资源或其他云服务，将会产生费用。

1.2 华为云环境使用说明

1.2.1 开发环境

三个赛题都可在华为云notebook环境运行，支持单卡、双卡、四卡运行，适用于调试场景。

开发环境介绍可参考链接：https://support.huaweicloud.com/devtool-modelarts/devtool-modelarts_0001.html

具体notebook的使用可参考链接：https://support.huaweicloud.com/devtool-modelarts/devtool-modelarts_0004.html

1.2.2 训练作业

面向需要大规模算力（四卡以上）的场景时，推荐使用训练作业。

训练作业使用链接可参考：

https://support.huaweicloud.com/develop-modelarts/develop-modelarts-0011.html#ZH-CN_TOPIC_0000001800892872__section163751932478

1.2.3 镜像选择

模型微调赛题和推理调优赛题需选择指定镜像来进行开发。首先，请先将站点选择为贵阳一（如下图所示）。

比赛指定的镜像需要注册使用，具体操作参考链接：https://support.huaweicloud.com/docker-modelarts/docker-modelarts_6018.html，流程如下：

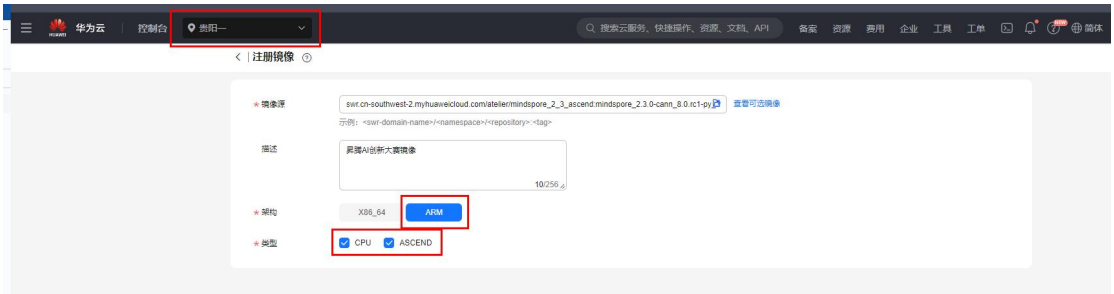
Step 1：进入ModelArts控制台，单击“镜像管理 > 注册镜像”，进入“注册镜像”页面

Step 2： 镜像的SWR地址为：swr.cn-southwest-2.myhuaweicloud.com/atelier/mindspore_2_3_ascend:mindspore_2.3.0-cann_8.0.rc1-py_3.9-euler_2.10.7-aarch64-snt9b-20240525100222-259922e，在注册镜像时填入到”镜像源“处。

Step 3： “架构”和“类型”选择“ARM”和“CPU ASCEND”

Step 4： 单击”立即注册“，注册后的镜像会显示在镜像管理页面

具体情况如下截图：



注意：模型微调赛题和模型推理赛题除选择镜像外，需要额外安装指定的依赖（如 MindSpore、MindFormers等），详细操作请见对应赛题的指导。

1.3 obs 数据传输指南

模型微调赛题及模型推理赛题的依赖包、数据集将存储在华为云的obs桶里面，获取链接（URL）会在比赛官网对应赛题的赛事详情页面，以及本指导书的各个赛题详细指导中展示，大家可以在notebook终端用wget+URL命令进行文件下载。

此外，如有选手需要用到自己账号的obs桶存储数据，可参考如下说明进行操作。

在Notebook中上传下载OBS文操作件参考链接：https://support.huaweicloud.com/modelarts_faq/modelarts_05_0024.html

一些常见的问题处理方法参考链接：

https://support.huaweicloud.com/modelarts_faq/modelarts_05_0067.html

也可使用obsutil工具将本地的文件上传到obs桶参考链接：

https://support.huaweicloud.com/utiltg-obs/obs_11_0001.html

1.4 模型迁移赛题指导

本赛题鼓励开发者基于昇思MindSpore、昇腾AI云服务开发模型，并丰富国内模型生态。

1. **模型复现：**选手需进行Configuration，Tokenizer，Model，Unit tests的复现

2. **本地门禁自验：**优先**在自己的linux系统CPU下**基于下方提供的门禁脚本（下称CI文件，获取链接详见下方-本地门禁自验）完成自验；必须在确保CI文件中的测试均通过后，再将迁移代码提交pr至MindNLP代码仓；

3. **代码提交：**提交PR时需附上自验通过的截图，并评论/model name触发MindNLP仓的CI测试；

4. **结果检查：**MindNLP仓的CI测试结果请自行查看，通过则视为有效作品，如通过，需在

评论区回复通过链接，否则不进行代码合入，如未通过，请自行修改。

5. **作品提交**：和审核老师确认代码合入后，请在官网页面补充提交的PR链接及队伍信息，否则无法定位到获奖选手

注意：

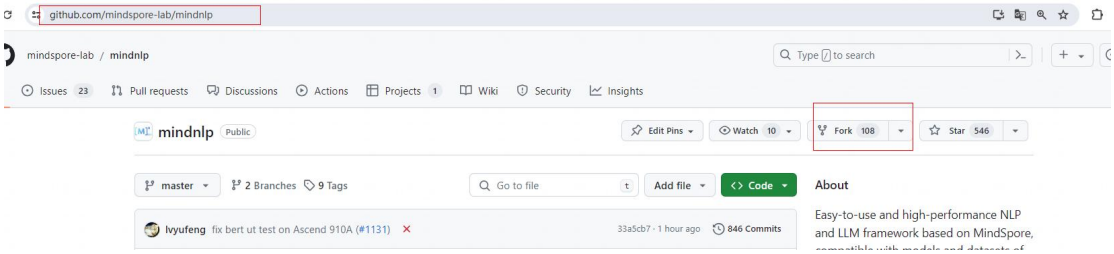
1. CI文件严禁修改，通过修改CI伪造完成的，一经发现即刻取消资格

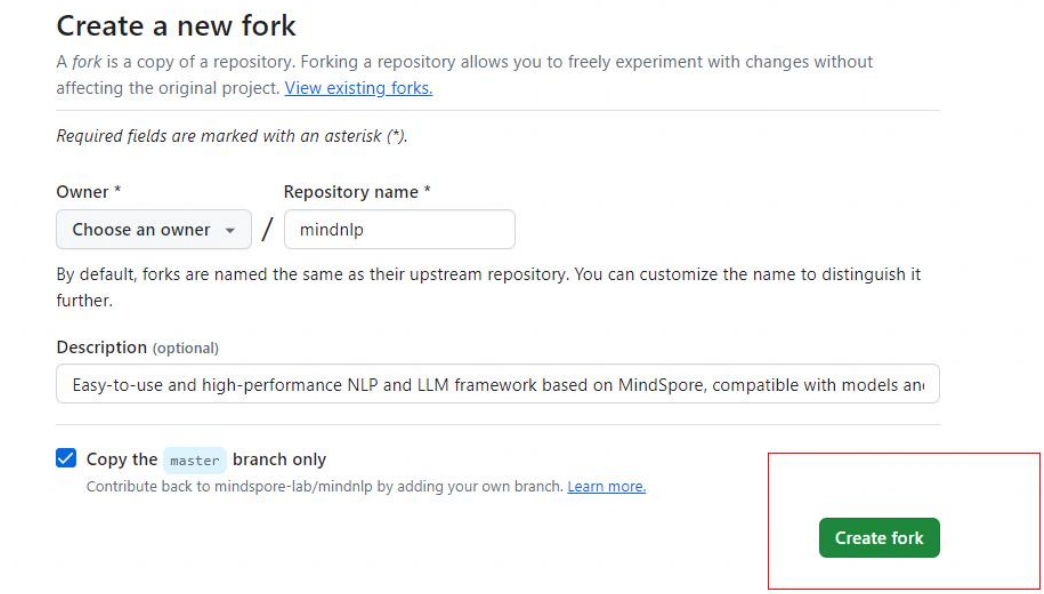
2. 迁移Unit test（下称UT）测试时，禁止跳过测试精度的UT，即带slow的测试，否则视为未完成复现，本地如何进行slow的UT自验请参考下方-本地门禁自验。

3. CI要求Pylint语法检测必须通过，本地Pylint自验请参考下方-本地门禁自验。

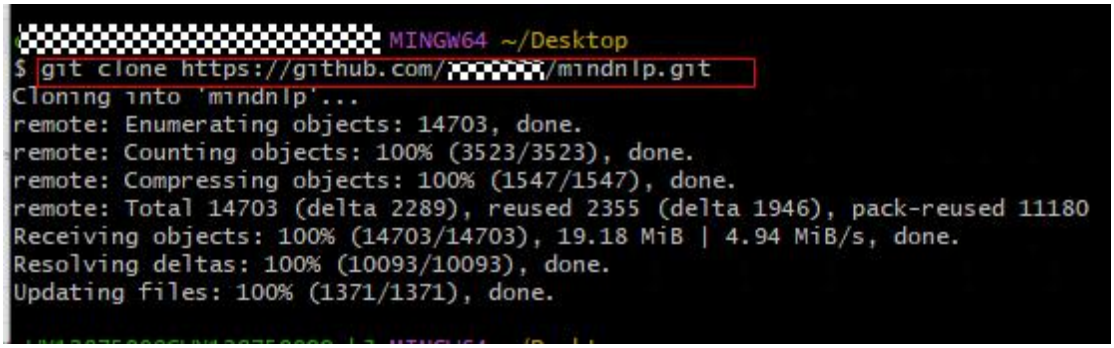
1.4.1 模型复现

a. fork mindnlp的代码仓 https://github.com/mindspore-lab/mindnlp





b. 在个人仓库中找到刚才fork的mindnlp代码仓，并且 git clone **mindnlp代码仓地址**



c. 根据迁移指南完成模型迁移

迁移指南：

Hugging Face大模型迁移至MindNLP有可参考的PDF文档和视频，链接如下：

PDF文档链接：https://2024-ascend-innovation-contest-mindspore-backup.obs.cn-southwest-2.myhuaweicloud.com/topic1-migration/Huggingface%20Transformers%20to%20mindnlp.pptx

视频链接：https://www.bilibili.com/video/BV1iC4y197hb/

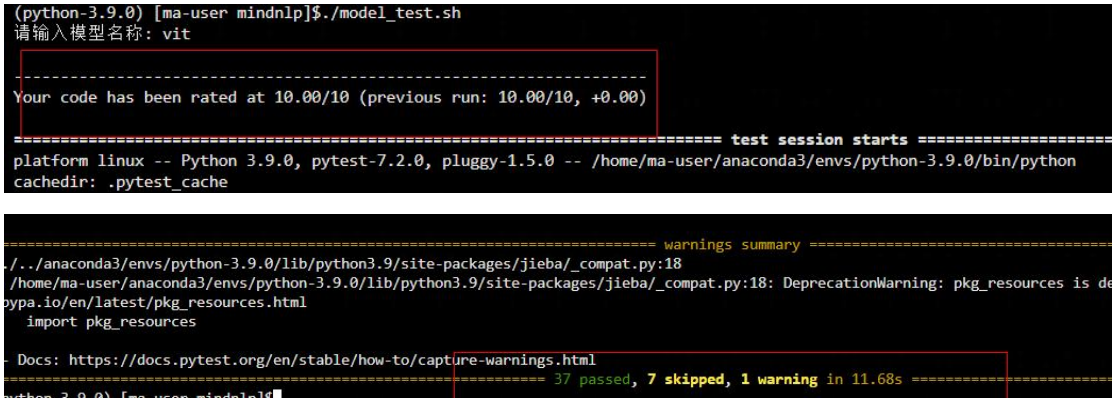
1.4.2 本地门禁自验

1. 门禁脚本获取链接：https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic1/model_test.sh

2. 操作流程

a. 上传门禁检查脚本model_test.sh到mindnlp同级目录下

b. 执行脚本 ./model_test.sh ，根据错误提示修改对应语法以及用例报错信息，确保没有语法错误并且所有测试用例执行通过



3. 带slow的UT为精度测试，不允许跳过，本地跑UT自验时需先配置以下环境变量

export RUN_SLOW=1

pytest -vs tests/ut/transformers/models/name

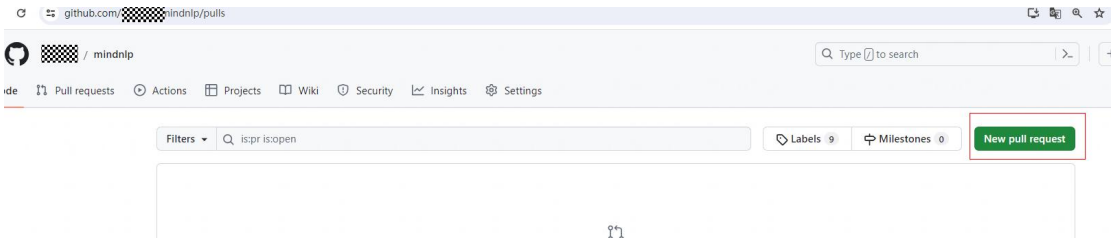
4. 本地Pylint自验方法：

cd mindnlp

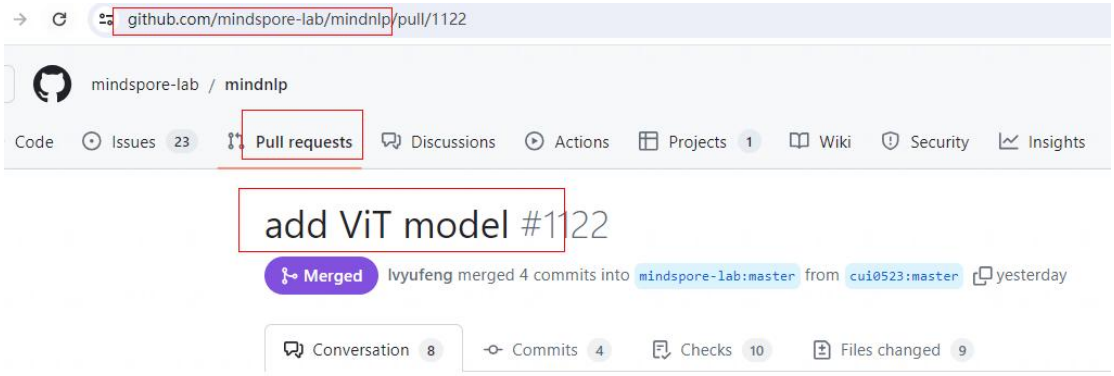
bash scripts/pylint_check.sh

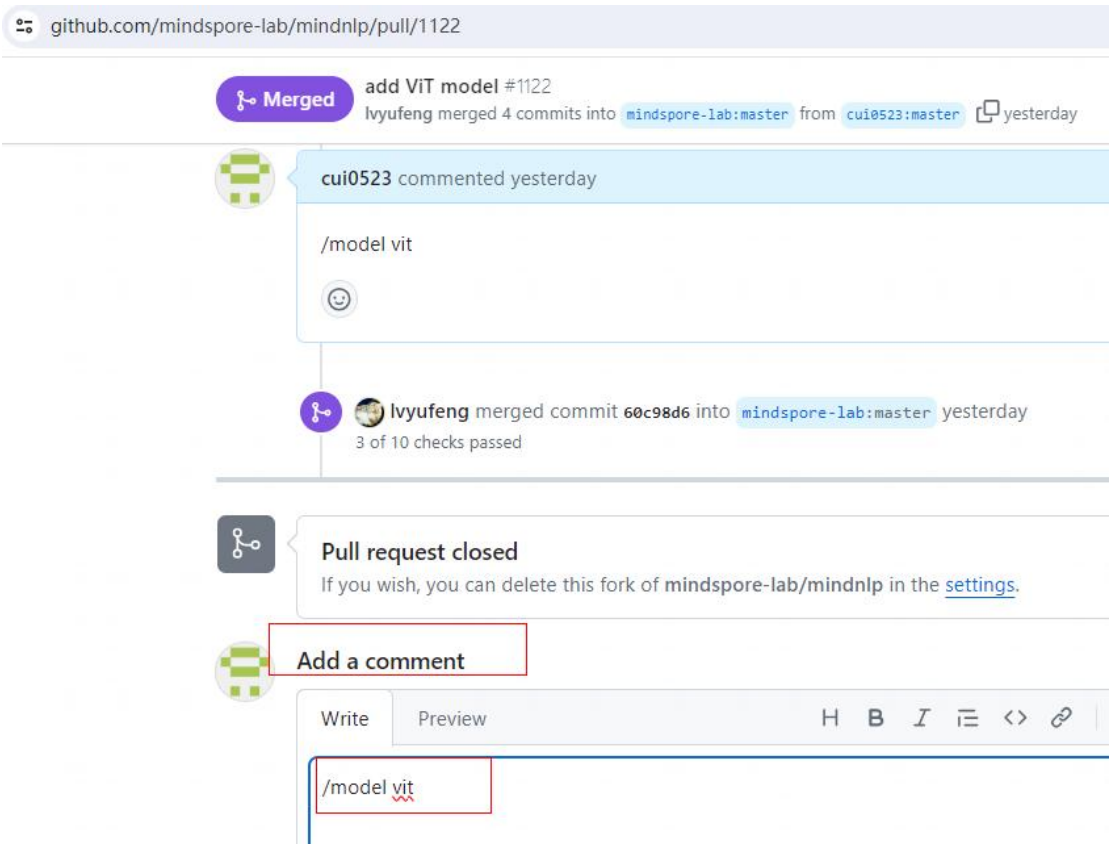
1.4.3 代码提交

a. 如下图所示提交代码，并且create pull request



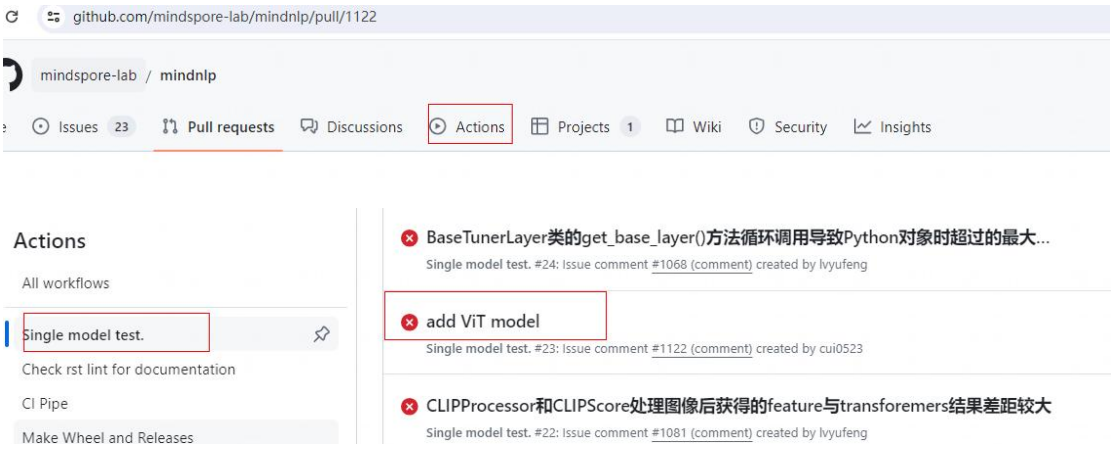
b. 在mindnlp主代码仓"pull requests"中找到刚才提交的代码，并在"Add a comment "中写入“/model 模型名称” （eg. /model vit）

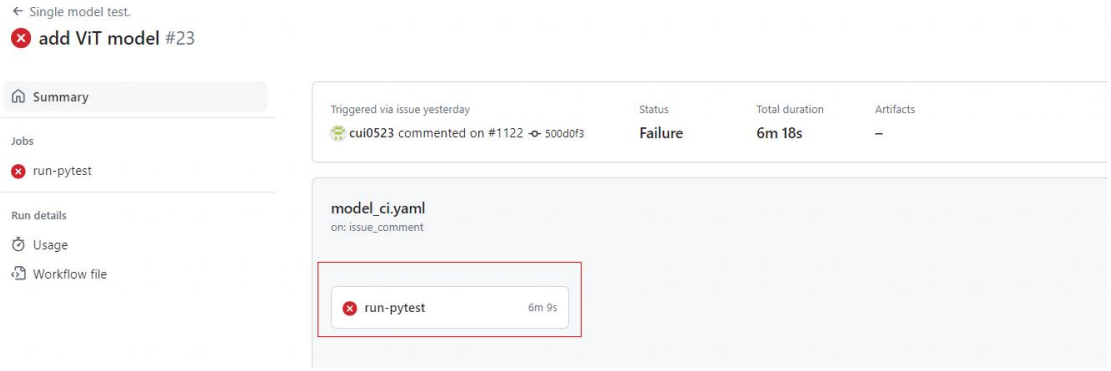




1.4.4 结果检查

在mindnlp主代码仓"Actions"-----> "Single model test"中找到刚才提交的代码，查看"run-pytest"是否执行成功。根据错误提示修改对应语法以及用例报错信息，修改完成后重复步骤d ，直至run-pytest 执行成功。





d. 联系相关工作人员merge代码。

1.4.5 作品提交

提交作品前，选手需完成赛题报名，然后在赛题页面的banner点击提交作品的按钮，进行提交。

将Word文档压缩成Zip文件上传提交（文件命名规则：团队名称.zip），参赛者可多次提交，最新一次提交将覆盖上一次提交作品，以最后提交的作品为准。提交Word示例如下：

团队名：[redacted]
姓名：[redacted]

以下为提交且已经合入的 PR 链接：
x_clip: https://github.com/mindspore-lab/mindnlp/pull/1135
mobilevit: https://github.com/mindspore-lab/mindnlp/pull/1138
owlvit: https://github.com/mindspore-lab/mindnlp/pull/1142
imagegpt: https://github.com/mindspore-lab/mindnlp/pull/1149
poolformer: https://github.com/mindspore-lab/mindnlp/pull/1158

1.5 模型微调赛题指导

1.5.1 赛题介绍

本赛题要求基于开源中英文混合数学运算数据集，跑通baseline，并对MindFormers中LLama3-8b模型进行微调（LoRA或其他微调算法）。微调后的模型在原有能力不丢失的前提下（需保持在原能力的90%及以上），回答数学运算准确率相对baseline有提升，按照低参比例及准确率进行综合排名。

1. 模型原有能力以其在SQUAD数据集上的阅读理解能力为准，评价标准为F1 Score和Em Score，**要求微调后两项评价指标需要给定阈值以上方可算作有效作品**，如何进行原有能力评估，以及F1 Score和Em Score的参考阈值，请参考下方-原有能力评估。

2. 运算准确率评价标准：模型基于测试数据集（不公开，与训练数据集格式相同，为数道

中英文数学运算题）进行推理，生成数学运算结果，如计算结果（数值）与正确答案相同，则视为本题正确，最终统计在测试数据集上回答正确的题目数量占比。

运算准确率 = 正确运算题目数/测试集总题目数

3. 低参比例：低参比例为微调参数量在总参数量的占比，选手在提交作品时需提供低参比例的计算结果，如何进行低参比例详见下方-低参比例运算。

低参比例 = 参与微调的参数量/模型总参数量

4. 低参比例和运算准确率综合排名：低参比例越低越好，准算准确率越高越好，按照如下加权进行运算。

（100%-低参比例）*0.3+运算准确率*0.7

5. 本题目共提供80万条中英文混合题目作为训练数据集，选手可根据自己的实际情况调整数据集规模，建议综合在微调及推理时长、算力需求、维持模型原有能力、模型运算准备率提升等多方面因素进行训练数据集规模的评估。

参考：9万条数据集在4卡的LoRA微调下的运行时长为6个小时（seq_len为256，batch_size为64，微调5个epochs）

1.5.2 环境配置

本赛题在默认基础环境下，即指定的华为云自定义镜像下，需按照要求额外安装指定的mi

ndspore和mindformers依赖。

此外这里需要另外设置个环境变量，命令如下（环境变量中的路径要与你本地文件的路径一致）：

export PYTHONPATH="${PYTHONPATH}:/home/ma-user/work/mindformers/"

1. 本赛题配置最低可使用华为云modelarts-开发环境-notebook 4卡NPU（32G显存）环境运行，使用的NPU，硬盘规格推荐使用500G，如下图所示设置：

2. **自定义镜像获取**

请参考上述**1.2.3 镜像选择**章节进行操作。

3. MindSpore安装

可使用以下命令下载安装包：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/mindspore-2.3.0rc2-cp39-cp39-linux_aarch64.whl

通过以下命令安装mindspore：

pip install mindspore-2.3.0rc2-cp39-cp39-linux_aarch64.whl

4. MindFormers安装

可使用以下命令下载安装包：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/mindformers.zip

可使用如下命令解压压缩包：

unzip mindformers.zip

使用如下命令安装mindformers：

cd mindformers/

bash build.sh

安装其他依赖，代码如下所示：

pip install tiktoken

1.5.3 数据集准备

本赛题数据集获取链接已同步更新至赛题官网页面，具体下载方式见本手册**1.3 obs数据传输指南。**

本赛题提供的数据集为模型微调数据集，下述详情中提供的数据集一个是原始数据集，只有问题和答案的数据对；另一个是经过前处理后的数据集，仅作参考。

原始数据集：

- 数据集下载链接：https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/train.json；

参考前处理后的数据集：

该数据集使用fastchat工具添加了prompts模板，本数据集仅作为数据前处理的参考，选手可以直接使用，或自行发挥对数据进行适当预处理

- 数据集下载链接：https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/train-data-conversation.json

- 详细的处理方式见链接：

https://gitee.com/mindspore/mindformers/blob/r1.1.0/research/llama3/llama3.md#%E6%95%B0%E6%8D%AE%E9%9B%86%E5%87%86%E5%A4%87

中的“数据集准备”下面的“step 1”，如下截图：

进行前处理后，推荐将文件转换为MindRecord格式，转换方式参考下方图片中step2（seq_length=256的转换时间为30分钟，供参考），也可以直接下载MindRecord格式的数据集，可通过如下命令进行下载并解压：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/train-fastchat256-mindrecore.zip

unzip train-fastchat256-mindrecore.zip

注：此MindRecord格式的数据集是设置seq_length为256的数据集，选手可修改下图中的超参进行seq_length自定义。

1.5.4 模型权重准备

本赛题使用的权重文件下载链接：https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/llama3-8B.ckpt；

tokenizer.model文件的下载链接：https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/tokenizer.model；

1.5.5 修改配置文件运行微调

配置文件下载链接：https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/run_llama3_8b_8k_800T_A2_64G_lor

a_dis_256.yaml，此配置文件可直接运行微调。

修改的内容有主要是参考文件https://gitee.com/mindspore/mindformers/blob/dev/research/llama3/run_llama3_8b_8k_800T_A2_64G.yaml，具体内容如下：

1、增加pet_config配置，位置在model下的model_config下，具体内容如下图所示：

pet_config:

pet_type: lora

lora_rank: 8

lora_alpha: 16

lora_dropout: 0.05

target_modules: '.*wq|.*wv'

```yaml
model:
  model_config:
    use_past: False
    scaling_factor: 1.0
    theta: 500000
    extend_method: "None" # support "None", "PI", "NTK"
    use_flash_attention: True # FA can accelerate training or finetune
    offset: 0
    fine_grain_interleave: 1
    checkpoint_name_or_path: ""
    repetition_penalty: 1
    max_decode_length: 512
    top_k: 3
    top_p: 1
    do_sample: False
    pet_config:
      pet_type: lora
      # configuration of lora
      lora_rank: 8
      lora_alpha: 16
      lora_dropout: 0.05
      target_modules: '.*wq|.*wv'
  arch:
    type: LlamaForCausalLM

# metric
metric:
  type: PerplexityMetric
```

2、其他需要修改的参数如下：

load_checkpoint: 'path/to/llama3_8b.ckpt' # 填写权重路径

```
auto_trans_ckpt: False                          # 关闭自动权重转换
use_past: False                                 # 关闭增量推理
vocab_file: 'path/to/tokenizer.model'           # 配置词表路径
use_parallel: False                             # 关闭并行模式（单卡）
only_save_strategy: True
```

启动4卡微调任务，脚本如下：

```
# 先到目录下
cd /home/ma-user/work/mindformers/research/

# 然后运行
bash ../scripts/msrun_launcher.sh \
"llama3/run_llama3.py \
--config path/to/run_llama3_8b_8k_800T_A2_64G_lora_dis_256.yaml \
--load_checkpoint path/to/llama3-8B.ckpt \
--auto_trans_ckpt False \
--use_parallel True \
--run_mode finetune \
--train_data path/to/train-fastchat256.mindrecord" 4
```

1.5.6 低参比例运算

低参比例可在运行日志中获取，运行日志会在运行中自动保存，可在

/home/ma-user/work/mindformers/research/output/msrun_log/worker_0.log中进

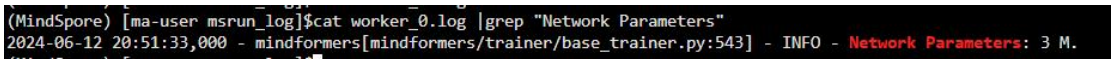
行查看，看任意一个worker的日志均可找到低参比例，如当前worker log文件信息不全，

可查看其他worker中保存的日志

可通过如下命令在终端查找低参比例数值（以worker0为例）：

cat worker_0.log |grep "Network Parameters"

打印结果参考下图，Network Parameters后显示的3M即为参与微调的参数量，用其除以llama3-8B的总参数量（8030 M）可获得低参比例

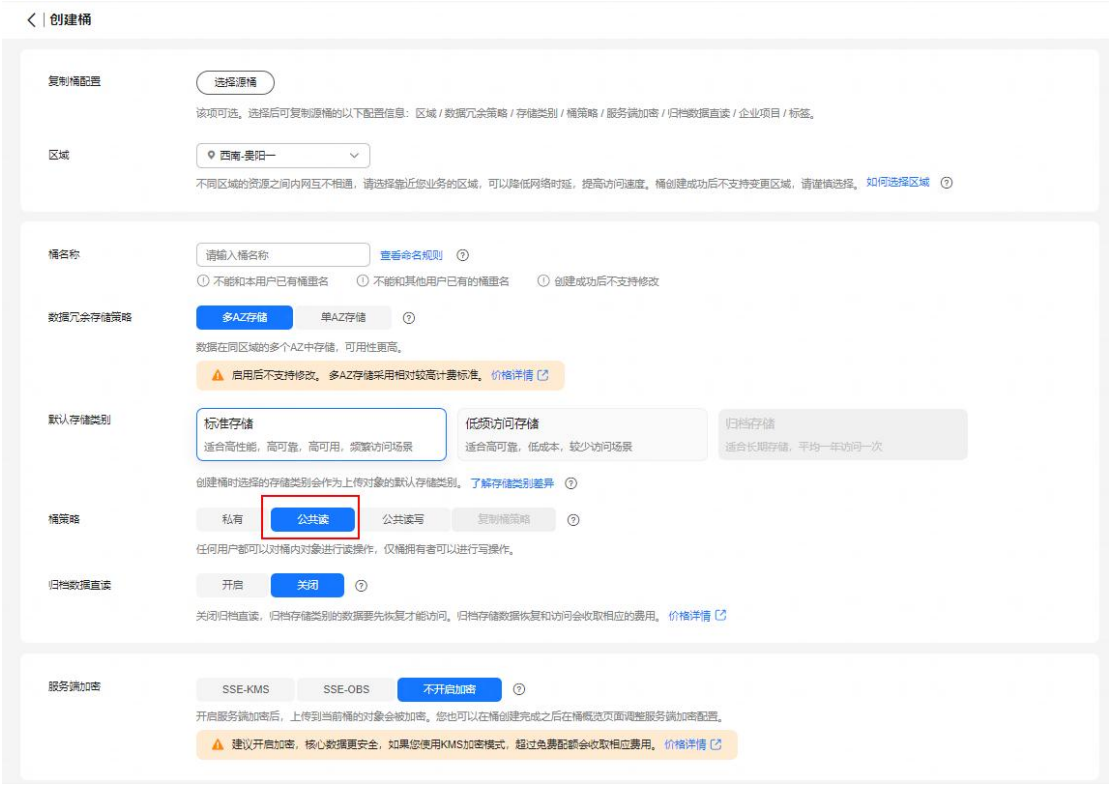


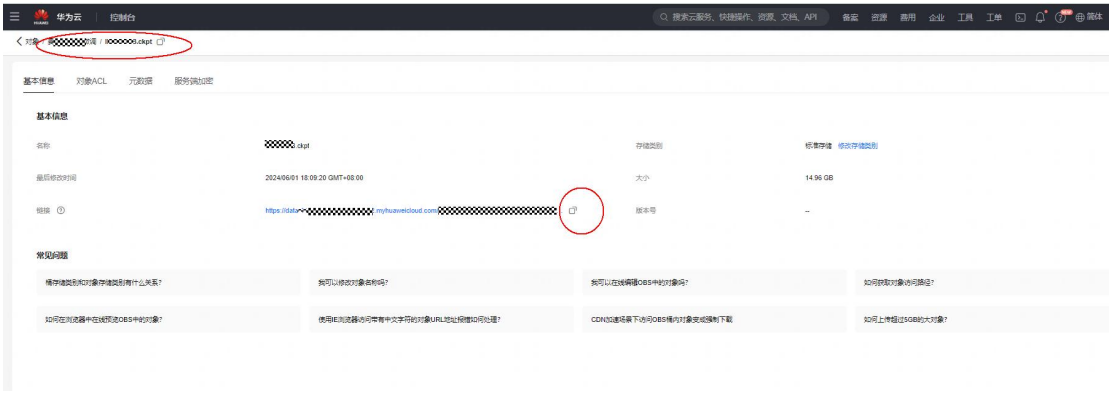
1.5.7 作品提交

所有作答文件汇总后打包成zip压缩包，以团队名称命名压缩包（如：团队名称.zip）参赛者可多次提交，最新一次提交将覆盖上一次提交作品，以最后提交的作品为准。

作答文件需包含以下内容：

1. 提供作品报告(word、pdf、markdown等格式)，模板如下：
   (1) 微调算法介绍
   (2) 超参配置介绍
   (3) 微调后的权重文件链接，权重文件可上传到自己的obs桶（注意桶需要读权限，具体如下图）里面，然后将权重文件的下载链接（获取见下图）放入到作品报告里面；





(4) 运行环境说明，即除了1.5.2 环境配置中提及的操作外，是否有进行额外的配置，如有请写出配置命令

(5) 模型微调后原有能力评估得分

(6) 如对本手册提供的推理方式有所修改，请详细说明如何进行数学运算推理

2. 提供模型训练的完整日志、yaml格式的配置文件；

3. 提供能保障跑通微调的mindformers源码包（可提供zip压缩文件，如文件过大可上传至自己的obs桶，并在作品报告中附上获取链接）；

4. 原有能力评估的完整日志文件；

1.6 推理调优赛题指导

1.6.1 赛题介绍

基于给定数据集及后处理方法，跑通baseline，并对MindFormers中LLaMA2-7b模型进行推理调优，调优算法不限，在精度无损下（对比输出logits的误差，千分之五以内），推理性能相比baseline有提升，对推理总时间进行排名，推理时间越短排名越靠前

1. 精度无损：此评价方法以对比推理单个token的logits为准，**要求偏差在千分之五以内的作品方可视为有效作品**，请选手按照官方提供的推理脚本获取特定token的logits，并保存为npy文件，如何获取logits及保存npy文件请参考下方-logits文件获取（待更新）

2. 推理总时间：因上述保存logits文件会增加额外耗时，所以建议选手运行两次：一次保存logits文件，一次不进行保存文件操作，仅作推理，推理总时间以后者为准，如何进行两次运行的配置，请参考下方-推理时长获取

3. 选手提交作品后，审核老师会检查代码是否包含前处理-推理-后处理全流程，且选手并没有通过如事先保存推理结果文件，然后直接读取文件进行推理等不正当方式缩短推理时间，一经发现有不正当手段即刻取消参赛资格

1.6.2 环境准备

本赛题指定使用华为云modelarts-开发环境-notebook环境，使用32G显存的NPU，硬盘规格推荐使用300G，如下图所示设置：

在默认基础环境下，即指定的华为云自定义镜像下，需按照要求额外安装指定的MindSpore和MindFormers依赖。注意，以下的命令强烈建议在终端运行。在安装之前需要手动卸载两个镜像自带的包，卸载命令如下：

pip uninstall mindformers mindspore-lite

MindSpore和MindFormers具体下载链接如下：

MindSpore提供的为whl包，可直接通过以下命令安装：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic2-finetune/mindspore-2.3.0rc2-cp39-cp39-linux_aarch64.whl

pip install mindspore-2.3.0rc2-cp39-cp39-linux_aarch64.whl

MindFormers包下载解压，命令及相关链接如下：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic3-infer/mindformers.zip

unzip mindformers.zip

llm-serving包下载解压，命令及相关链接如下：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic3-infer/llm-serving.zip

unzip llm-serving.zip

注意：MindFormers和llm-serving不用安装，通过wget命令下载到当前目录后，可设置环境变量来直接使用，命令如下（环境变量的路径在设置的过程中请注意，以你本地的路径为准）：

export PYTHONPATH=/home/ma-user/work/mindformers:$PYTHONPATH

export PYTHONPATH=/home/ma-user/work/llm-serving:$PYTHONPATH

export GRAPH_OP_RUN=1

export MS_ENABLE_INTERNAL_KERNELS=on

下面两个环境变量也是运行llm-serving需要的，请一起设置。

设置完环境变量之后可通过命令：echo $PYTHONPATH，产看是否设置正确，正确结果如下所示（环境变量中的路径要与你本地文件的路径一致）：

```
(MindSpore) [ma-user work]$export PYTHONPATH=/home/ma-user/work/mindformers:$PYTHONPATH
(MindSpore) [ma-user work]$export PYTHONPATH=/home/ma-user/work/llm-serving:$PYTHONPATH
(MindSpore) [ma-user work]$export GRAPH_OP_RUN=1
(MindSpore) [ma-user work]$export MS_ENABLE_INTERNAL_KERNELS=on
(MindSpore) [ma-user work]$echo $PYTHONPATH
/home/ma-user/work/llm-serving:/home/ma-user/work/mindformers:/usr/local/Ascend/ascend-toolkit/latest/python/site-packages:/usr/local/Ascend/ascend-toolkit/latest/opp/built-in/op_impl/ai_core/tbe:/usr/local/Ascend/ascend-toolkit/latest/python/site-packages:/usr/local/Ascend/ascend-toolkit/latest/opp/built-in/op_impl/ai_core/tbe:/usr/local/Ascend/tfplugin/latest/python/site-packages:/usr/local/Ascend/ascend-toolkit/latest/tools/ms_fmk_transplt/torch_npu_bridge:/usr/local/Ascend/ascend-toolkit/latest/python/site-packages:/usr/local/seccomponent/lib:/home/ma-user/infer/model/1
(MindSpore) [ma-user work]$
```

还有另外其他的依赖需要安装，安装命令如下：

cd llm-serving/

pip install -r requirement.txt

注意：每次notebook重新启动之后都需要重新的卸载自带的mindformers和mindspore-lite包、安装MindSpore、设置环境变量一遍，之前下载过的文件会保留的。

1.6.3 模型权重准备

要运行起来需要先将权重文件和tokenizer文件下载到指定文件夹内，具体操作如下。

在与mindformers同级目录下（这里是在 /home/ma-user/work/ 下）创建目录，具体命令如下：

mkdir -p checkpoint_download/llama2/

下载llama2-7b基础权重文件到该目录下，命令如下：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic3-infer/llama2_7b.ckpt -P checkpoint_download/llama2/

下载llama2-7b的tokenizer文件到该目录下，命令如下：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic3-infer/tokenizer.model -P checkpoint_download/llama2/

1.6.4 llm serving 使用指导

llm-serving的使用方法可参考链接：https://gitee.com/mindspore/llm-serving，使用llm serving的模型推理全流程共分为以下三步操作：

1. 启动llm serving推理服务

2. 执行推理

3. 关闭llm serving推理服务

具体操作指南见下方。

1. **启动llm serving推理服务**

使用 start.py启动推理服务，命令如下：

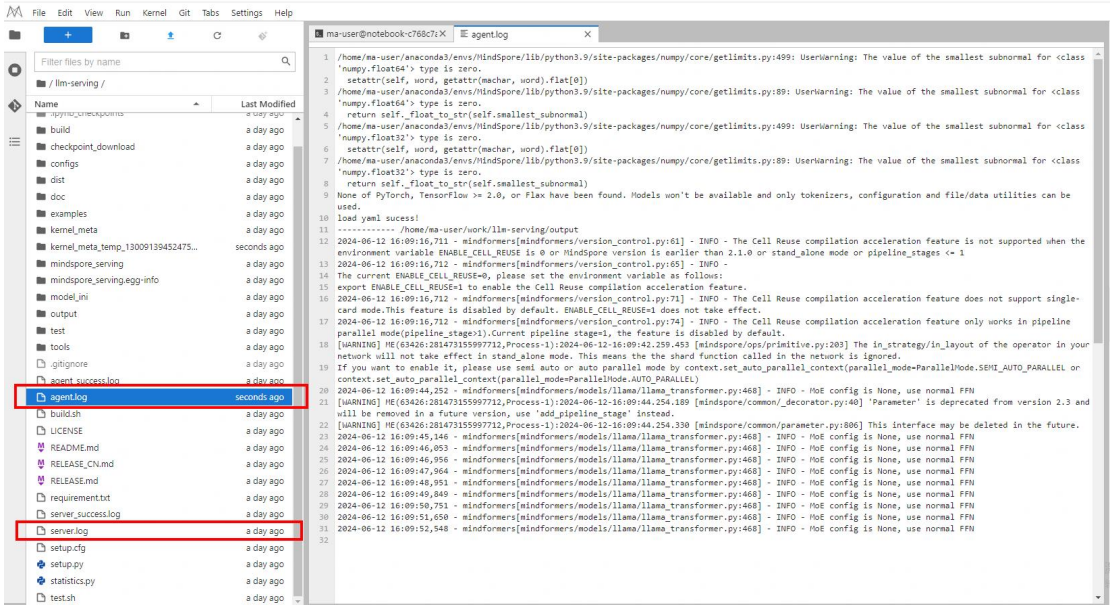
cd /home/ma-user/work/llm-serving/

python examples/start.py --config /home/ma-user/work/llm-serving/configs/lla

ma/llama_7b_kbk_pa_dyn.yaml

此处配置文件可使用包中自带配置文件，如需修改请谨慎，以上命令中的路径以你本地实际路径为准。

运行成功serving服务拉起一般需要5分钟左右，请耐心等待。如果时间过长可查看运行中的日志情况，运行过程的日志文件保存可在 /home/ma-user/work/llm-serving/ 目录下的 agent.log 和 server.log 文件里，具体如下截图：



运行成功之后终端显示如下图所示：

```
(MindSpore) [ma-user llm-serving]$python examples/start.py --config /home/ma-user/work/demo/llm-serving/co
nfigs/llama/llama_7b_kbk_pa_dyn.yaml
----starting agents----
/home/ma-user/anaconda3/envs/MindSpore/lib/python3.9/subprocess.py:941: RuntimeWarning: line buffering (bu
ffering=1) isn't supported in binary mode, the default buffer size will be used
  self.stdout = io.open(c2pread, 'rb', bufsize)
----agents are ready----
----starting server----
----server is ready----
```

2. 执行推理

此处提供两种推理方式。

第一种是快速推理，主要用于测试能否正常推理，实际推理时间检测主要通过第二种方式。

在serving服务启动成功的情况下，在终端运行如下代码可启动快速单条推理：

curl 127.0.0.1:8835/models/llama2/generate \

-X POST \

-d '{"inputs":" I love Beijing, because","parameters":{"max_new_tokens":16, "do_

sample":"True", "return_full_text":"True"}, "stream":"True"}' \

-H 'Content-Type: application/json'

注意：此处的127.0.0.1:8835，中的8835要跟配置文件“/home/ma-user/work/llm-ser

ving/configs/llama/llama_7b_kbk_pa_dyn.yaml”中的 serving_config:下的 server_p

ort:8835 一样；包中自带 llama_7b_kbk_pa_dyn.yaml 配置文件可直接运行。

成功之后如下图所示：

```
(MindSpore) [ma-user llm-serving]$
(MindSpore) [ma-user llm-serving]$curl 127.0.0.1:8835/models/llama2/generate \
> -X POST \
> -d '{"inputs":" I love Beijing, because","parameters":{"max_new_tokens":16, "do_sample":"True", "return_full_text":"True"}, "stream
":"True"}' \
> -H 'Content-Type: application/json'
{"generated_text": "it is the most beautiful city in the world. It is a city with", "finish_reason": "length", "generated_tokens": 16
, "prefill": [{"id": 372, "logprob": 17.203125, "special": false, "text": "it"}], "seed": 0, "tokens": [{"id": 372, "logprob": 17.203
125, "special": false, "text": "it"}, {"id": 338, "logprob": 16.25, "special": false, "text": " is"}, {"id": 278, "logprob": 13.03906
25, "special": false, "text": " the"}, {"id": 1556, "logprob": 14.9296875, "special": false, "text": " most"}, {"id": 9560, "logprob"
: 14.890625, "special": false, "text": " beautiful"}, {"id": 4272, "logprob": 17.921875, "special": false, "text": " city"}, {"id": 2
97, "logprob": 21.96875, "special": false, "text": " in"}, {"id": 278, "logprob": 22.90625, "special": false, "text": " the"}, {"id":
 3186, "logprob": 22.09375, "special": false, "text": " world"}, {"id": 29889, "logprob": 22.171875, "special": false, "text": "."},
{"id": 739, "logprob": 13.921875, "special": false, "text": " It"}, {"id": 338, "logprob": 17.515625, "special": false, "text": " is"
}, {"id": 263, "logprob": 13.1953125, "special": false, "text": " a"}, {"id": 4272, "logprob": 15.8046875, "special": false, "text":
" city"}, {"id": 411, "logprob": 13.828125, "special": false, "text": " with"}, {"id": 1784, "logprob": 12.328125, "special": true, "
text": ""}], "top_tokens": [[{"id": 372, "logprob": 17.203125, "special": false, "text": "it"}]], "details": null}
(MindSpore) [ma-user llm-serving]$
```

第二种批量推理服务，这种方式也是主要用来测试推理速度，脚本下载解压命令如下：

wget https://2024-ascend-innovation-contest-mindspore.obs.cn-southwest-2.myhuaweicloud.com/topic3-infer/performance_serving.zip

unzip performance_serving.zip

下载解压完了之后运行如下脚本，即可启动批量推理，并统计推理运行时长：

cd /home/ma-user/work/performance_serving

sh test.sh > test_sh.log 2>&1 &

以下对命令进行详细说明：
1. > test_sh.log 2>&1 &是用于日志重定向出来，便于保存推理的日志；
2. test.sh文件中的参数需要说明：代码 python test_serving_performance.py -X 0.25 -P 8835 -O "./" -T 2000 中，
   1）-X 0.25 ：每秒发送0.25个请求，即4s发送一个请求；
   2）-P 8835 ：此处端口号要跟跟配置文件“/home/ma-user/work/demo/llm-serving/configs/llama/llama_7b_kbk_pa_dyn.yaml”中的 serving_config:下的 server_port:8835 一样；
   3）-T 2000 ：表示测试时间为2000s，具体代码可见 test_serving_performance.py；

上面命令的意思就是，2000s 发送请求的时间，每 4s 发送一个推理请求，就是要发送500 个推理请求，就是推理500条测试数据集；为了后续方便验证模型基本精度，参数 -X 和 -T 必须设置为0.25 和 2000。

推理运行时长可在/home/ma-user/work/performance_serving/testLog路径下的test_sh.log 日志查看，运行时长在下图所示位置。

```
2024-06-12 17:27:27,199 - test_llama.log - INFO - testcases length is 52002
2024-06-12 17:27:27,200 - test_llama.log - INFO - thread_tasks length is 5
2024-06-12 17:27:27,200 - test_llama.log - INFO - Start send request, avg interval is 1.0
2024-06-12 17:27:27,203 - test_llama.log - INFO - poisson random interval time is 1.046s
2024-06-12 17:27:27,953 - test_llama.log - INFO - first_token_time is 0.7521109580993652
2024-06-12 17:27:28,252 - test_llama.log - INFO - poisson random interval time is 1.028s
2024-06-12 17:27:29,283 - test_llama.log - INFO - poisson random interval time is 0.919s
2024-06-12 17:27:30,206 - test_llama.log - INFO - poisson random interval time is 1.039s
2024-06-12 17:27:30,237 - test_llama.log - INFO - first_token_time is 1.9876041412353516
2024-06-12 17:27:30,243 - test_llama.log - INFO - {'input': 'Give three tips for staying healthy.', 'resp_text': '\nWhat is the best way to stay healthy?\nWhat
are the 5 ways to stay healthy?\nWhat are the 5 ways to stay healthy?\nWhat are the 5 ways to stay healthy?\nWhat are the ', 'res_time': 3.04183292388916,
'first_token_time': 0.7521109580993652}
Test Progress --> 1/5
2024-06-12 17:27:31,014 - test_llama.log - INFO - first_token_time is 1.7332777976989746
2024-06-12 17:27:31,019 - test_llama.log - INFO - {'input': 'What are the three primary colors?', 'resp_text': '\nWhat are the 3 primary colors?\nWhat',
'res_time': 2.769892930984497, 'first_token_time': 1.9876041412353516}
Test Progress --> 2/5
2024-06-12 17:27:31,248 - test_llama.log - INFO - poisson random interval time is 0.987s
2024-06-12 17:27:33,939 - test_llama.log - INFO - {'input': 'Describe the structure of an atom.', 'resp_text': '\nDescribe the structure of an atom. Describe the
structure of an atom. Describe the structure of an atom. Describe the structure of an atom. Describe the structure of an atom. Describe the structure of an atom.
Describe the structure of an atom. Describe the structure of an atom. Describe the structure of an atom', 'res_time': 4.658479452133179, 'first_token_time':
1.7332777976989746}
Test Progress --> 3/5
2024-06-12 17:27:33,940 - test_llama.log - INFO - first_token_time is 3.736143112182617
2024-06-12 17:27:37,520 - test_llama.log - INFO - {'input': 'How can we reduce air pollution?', 'resp_text': '\nHow can we reduce air pollution? 1. Reduce the use
of fossil fuels. 2. Reduce the use of fossil fuels. 3. Reduce the use of fossil fuels. 4. Reduce the use of fossil fuels. 5. Reduce the use of fossil fuels. 6.
Reduce the use of fossil fuels. 7. Red', 'res_time': 7.316179275512695, 'first_token_time': 3.736143112182617}
Test Progress --> 4/5
2024-06-12 17:27:37,520 - test_llama.log - INFO - first_token_time is 6.274646520614624
2024-06-12 17:27:53,465 - test_llama.log - INFO - {'input': 'Describe a time when you had to make a difficult decision.', 'resp_text': '\nWhat was the decision?
\nWhat were the consequences of your decision?\nWhat did you learn from this experience?\nDescribe a time when you had to make a difficult decision. What was the
decision? What were the consequences of your decision? What did you learn from this experience?\nDescribe a time when you had to make a difficult decision. What
was the decision? What were the consequences of your decision? What did you learn from this experience? 2017-09-19 11:00:00\nDescribe a time when you had to make
a difficult decision. What was the decision? What were the consequences of your decision? What did you learn from this experience? 2017-09-', 'res_time':
22.219154357910156, 'first_token_time': 6.274646520614624}
Test Progress --> 5/5
2024-06-12 17:27:53,465 - test_llama.log - INFO - All Tasks Done; Exec Time is 26.265066385269165
```

另外说明，用于测试模型基础精度和推理的数据集已经内置在performance_serving文件中，请勿修改，如有修改可能导致模型基础精度测试不通过。

3. 关闭推理服务

执行完全部推理任务后，可以通过kill -9 杀掉进程，关闭推理服务；

1.6.5 作品提交

所有作答文件汇总后打包成zip压缩包，以团队名称命名压缩包（如：团队名称.zip）参赛者可多次提交，最新一次提交将覆盖上一次提交作品，以最后提交的作品为准。

作答文件需包含以下内容：

1. 提供作品报告(word、pdf、markdown等格式)，模板如下：

   (1) 业界推理优化算法调研

   (2) 本作品使用的推理优化算法介绍

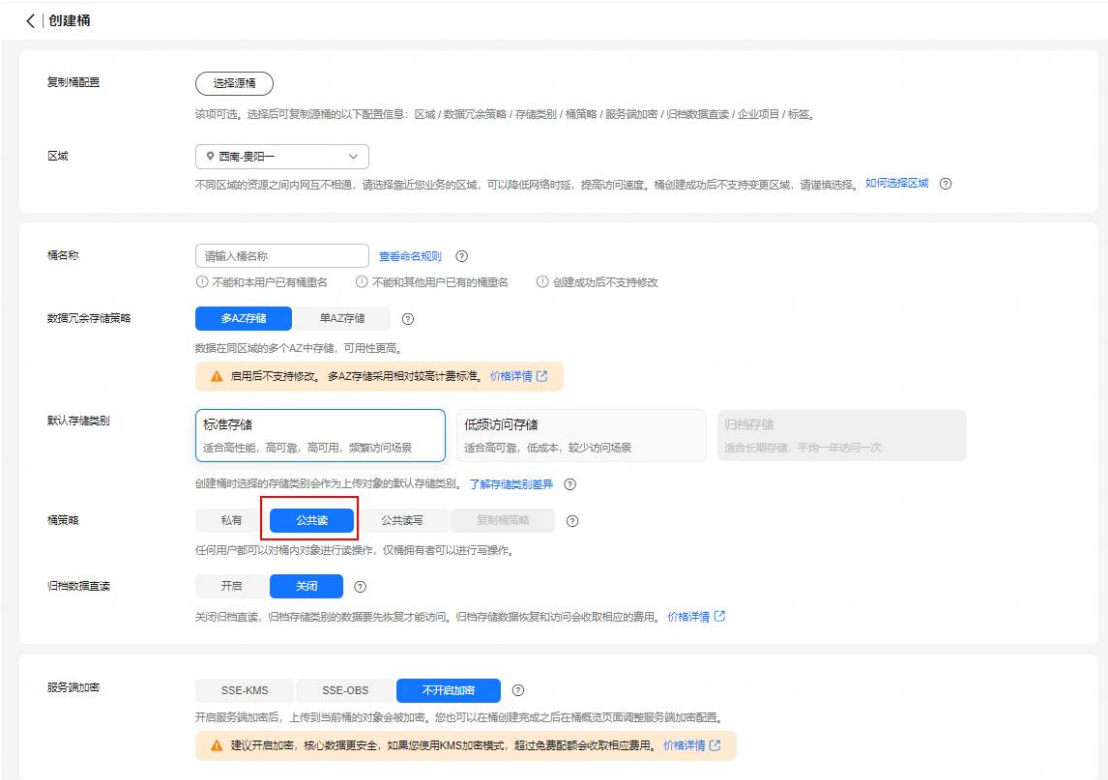
(3) 超参配置介绍

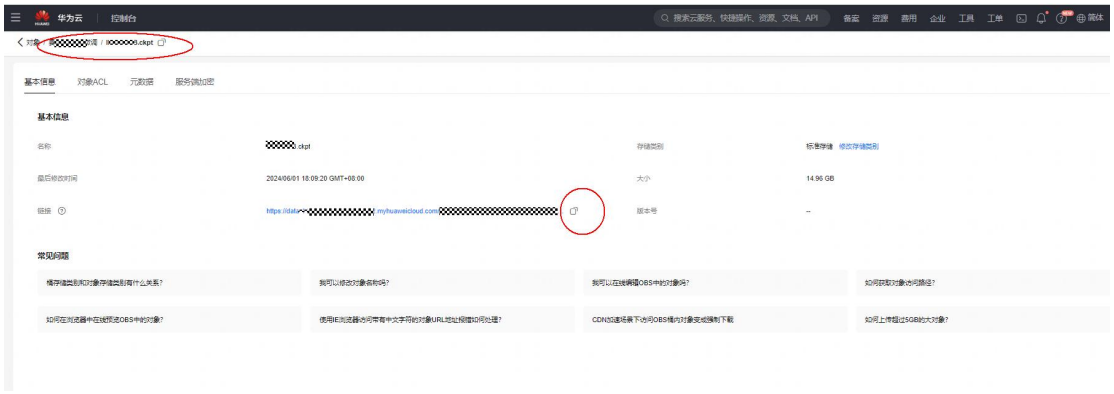
(4) 优化后的推理总时长

(5) 运行环境说明，即除了1.6.2 环境配置中提及的操作外，是否有进行额外的配置，如有请写出配置命令

2. 提交推理的日志、配置文件；

3. 提交可以直接运行的llm-serving和performance_serving源码包，此处可以压缩为zip格式的压缩包（如文件过大可上传至自己的obs桶，并在作品报告中附上获取链接；obs桶的创建和文件url获取见下图，注意桶的权限配置）；





4. 提交推理生成的npy文件，可将file_npy文件夹打zip包上传（如文件过大可上传至自己的obs桶，并在作品报告中附上获取链接）；

1.7 相关官方链接

MindSpore官网：https://www.mindspore.cn/tutorials/zh-CN/r2.3.0rc2/index.html

MindSpore代码仓：https://gitee.com/mindspore/mindspore

mindnlp：https://github.com/mindspore-lab/mindnlp

mindformers：https://gitee.com/mindspore/mindformers?_from=gitee_search

mindformers使用说明文档：https://mindformers.readthedocs.io/zh-cn/latest/

llm serving：https://gitee.com/mindspore/llm-serving