

12/15 meeting

Friday, December 15, 2023 10:55 AM

- 1.工作内容了解: PVC(intel gpu)与a100的对比, 微调速度
- 2.transformers版本问题, gpt-j只能到4.26.最新4.36, 优先支持llama2等
Gpt-j 在cpu升级4.36只能一个worker运行
Gpu >= 4.36
- 3.

Monday, December 18, 2023 1:10 PM

```
source$(python -c "import oneccl_bindings_for_pytorch as torch_ccl;print(torch_ccl.cwd)"/env/setvars.sh
```

Start the Ray head node using the following command.

更改trainer tmp位置

2.非gpt-base模型需要更改

3.ray的tmp会布置临时运行环境, 添加transformers, 来应对不同模型的需求

Gpt-base需要transformers==4.26.0

其他模型用最新，和conda env保持一致

model	finetune-data	Cpu core	Num-training-workers	loss	iteration-time	Epoch/step
Llama-2-7b-chat-hf	examples/data/sample_finetune_data_small.jsonl	32	2		10min-16min	3/7
Gpt2	examples/data/sample_finetune_data_small.jsonl					
Gpt-j-6b	examples/data/sample_finetune_data_small.jsonl					

Memory used more than 50%

Cpu used 5% sometime

5.模型serving情况

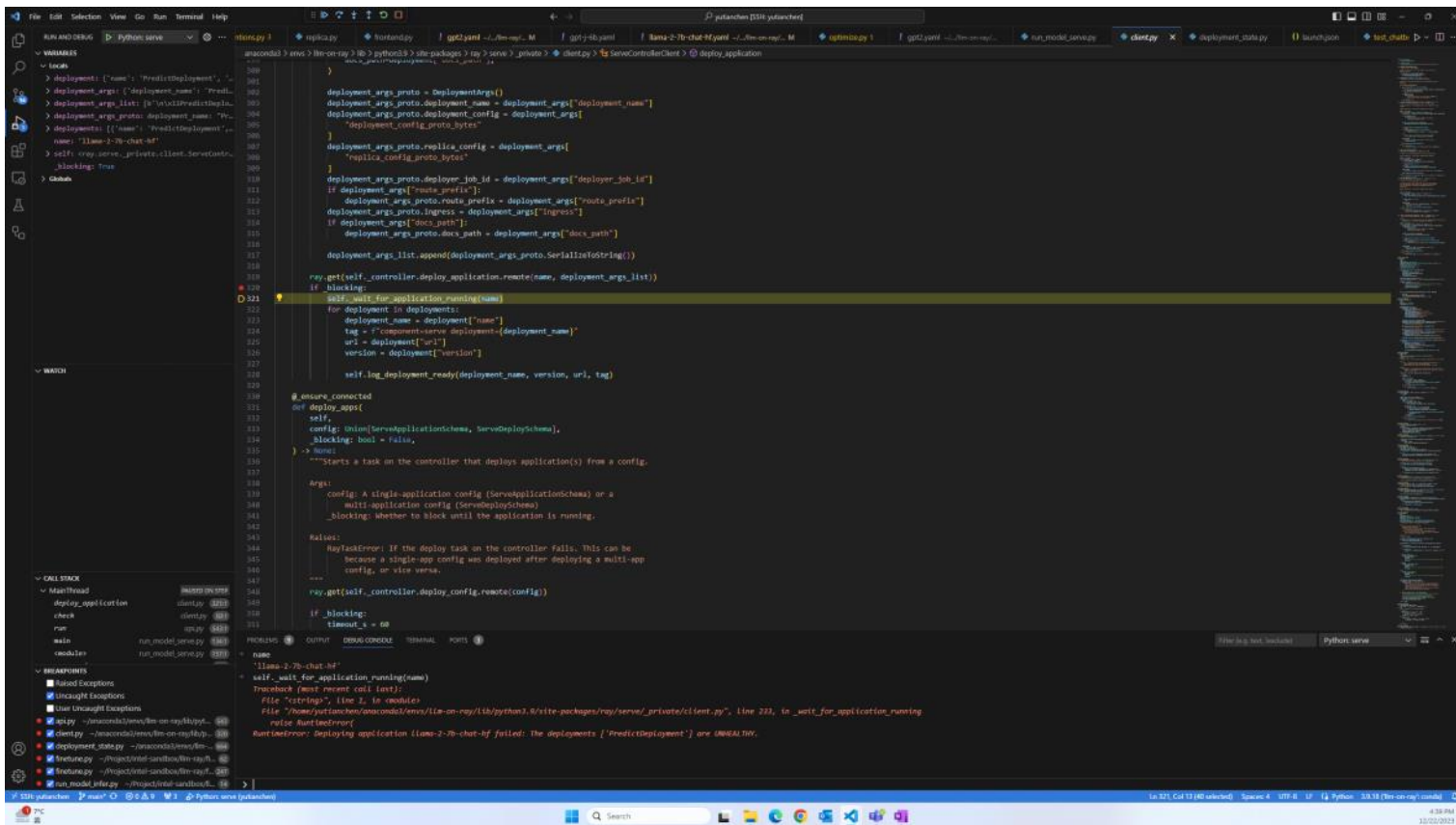
llama运行, Transformers 版本太高?

降低到4.31.0

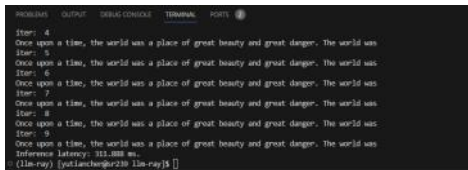
降至4.35.2 success

Ln 35, Col 11B (16 selected) Spc

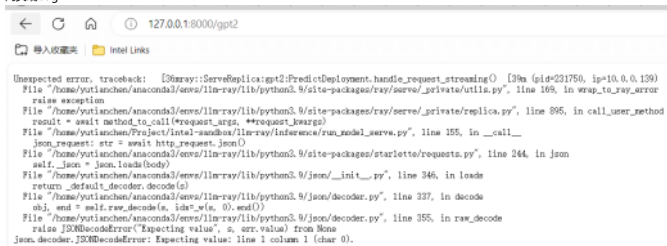
出现nn层无法运行的bug



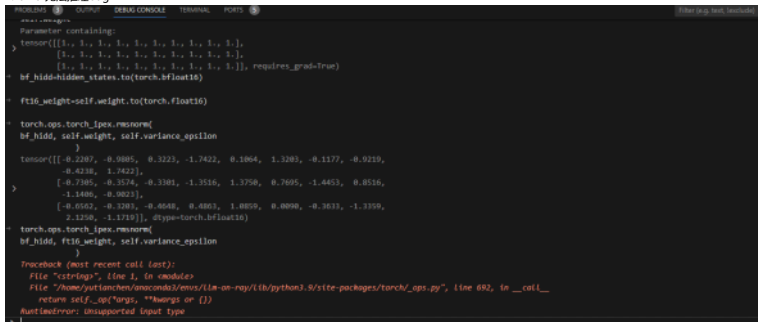
Gpt2运行, success



网页端bug



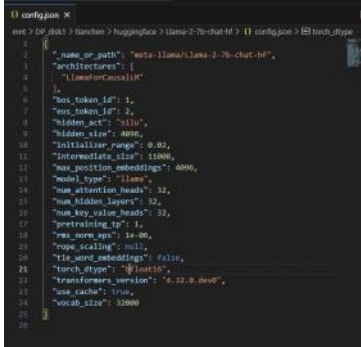
7llama 无法推理bug



hiddenstates和weight不相同

[Issue about using torch.onnx + Issue 2197 + Intel/llm-on-ray \(github.com\)](#)

cpu不支持float16的推理, 需要将llama的config文件改为bf16float16



Llama 推理速度: 任务 once opinion a time续写, 不设置token上限

CPU	42	56	112*?
单个iter时间	100s	80s	80s

6.deepspeed
Worker per group =1
注意onecl 环境变量

是否用deepspeed速度对比

Transformers=4.35.2, 10轮文本生成均值

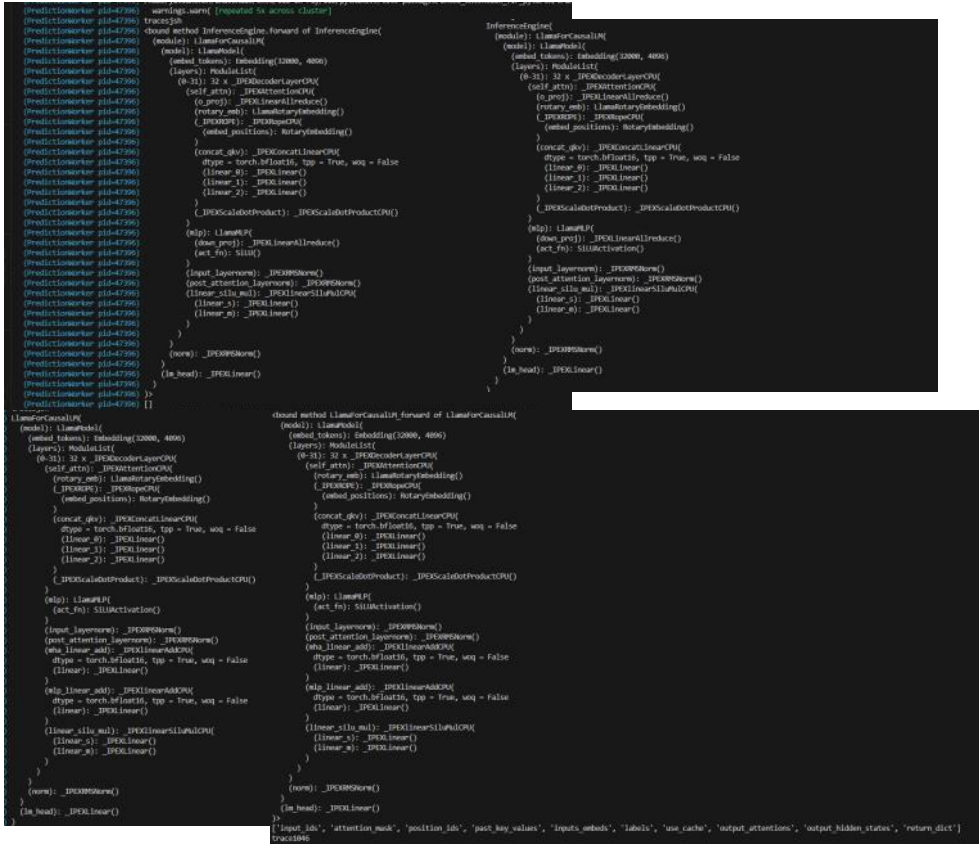
模型	Cpu/ num	Max_token	cpu	Deepspeed+cpu
Gpt2	24/1	None	837ms	590ms
Llama2-7b	24/1	None	97368ms(ipex true) 176291ms(ipex false)	85522ms(ipex true) 105726ms (ipex false)
Gpt-j-6b	4.35.2不支持	none		
Bloom-560m	10/1	none	850ms (ipex true)	1007ms(ipex true)
Llama2-7b	28/1	128	26654ms(ipex true) 50189ms(ipex false)	25000ms(ipex true) 30000ms(ipex=false 注释module)
Llama2-7b	28/2	128	26000ms(ipex true) 45632ms(ipex false)	24500ms(ipex true) 27000ms(ipex=false 注释module)
Mistral-7b	24/2	128	5318ms(ipex true) 38305ms(ipex false)	34518ms(ipex true) 47026ms(ipex false)
neural	24/2			25984ms(ipex true)

Deepspeed forward method

[some docs for ci facilities](#) * [intel-sandbox/llm-ray@87b73b](#) ([github.com](#)).

llama的ipex enable 定为false, deepspeed 定为true

Deepspeed 将model套了一层interface, 需要model=model.module操作



Pytest-llm

Tuesday, December 19, 2023 9:19 AM

- 1.[pytest: helps you write better programs — pytest documentation](#)
- 2.[ray-llm/tests at master · ray-project/ray-llm \(github.com\)](#)
- 3.[confident-ai/deepeval: The Evaluation Framework for LLMs \(github.com\)](#)

Test 计划 从inference开始，对每个py文件，稍微复杂的

第一步搭建测试框架

周五讨论

Pytest---如何使用--capture参数设置实时打印及Capture Log打印



DevOps技术社区 | 2022-08-28 11:28:03

目录

- 一、设置标准输出标准错误输出的模式
 - 1.1 pytest命令
 - 1.2 pytest -s 命令
 - 1.3 pytest --capture=sys 命令
 - 1.4 pytest --capture=fd 命令
 - 1.5 pytest --capture=tee-sys 命令
- 二、在测试函数中捕获标准输出标准错误输出

一、设置标准输出标准错误输出的模式

pytest捕获标准输出标准错误输出的模式主要有以下几种

- pytest：和pytest --capture=fd 模式是一样的，默认的就是pytest --capture=fd 模式

proxy问题

Tuesday, December 19, 2023 3:18 PM

1.huggingface proxy terminal input

```
export https_proxy="http://10.24.221.149:911"
```

```
export http_proxy="http://10.24.221.149:911"
```

```
git config --global --edit
```

来自 <<https://zhuanlan.zhihu.com/p/390874935>>

2.intel github proxy

```
dt proxy print
```

From <https://1source.intel.com/docs/getting_started/environment_setup#manual-configuration>

3.intel proxy

child-prc.intel.com:912

[linux设置代理教程 - 知乎 \(zhihu.com\)](#)

4.wget proxy

<http://proxy-shz.intel.com:912>

加入sudoers后就不能通过这个proxy下载了

5.docker proxy

[Docker的三种网络代理配置 · 零壹轩 · 笔记 \(qidong.name\)](#)

6.windows pip 需要intel 代理

Git+scp

Friday, December 22, 2023 2:26 PM

2、pull拉取最新代码

本地仓库的代码还未被更新，此时：

(1) 更新远程仓库的代码为最新的

```
1 | git fetch --all
```

(2) 让本地代码与origin / master完全相同

```
1 | git reset --hard origin/master
```

(3) git pull拉取远程代码

```
1 | git pull origin master
```

(4) git merge 将暂存区代码更新到本地工作区

```
1 | git merge master
```

[Git克隆、拉取最新代码、上传代码到远程仓库 git pull拉取代码-CSDN博客](#)

下载huggingface

Git lfs install --skip-smudge

Git clone http。。。

Cd

Git lfs pull

```
scp -o "ProxyCommand ncat --proxy-type=socks5 --proxy 10.239.44.250:1080 %h %p" .\scptest.txt  
root@10.0.0.139:/home/yutianchen/Project/
```

Git pr ci 流程

Friday, January 5, 2024 11:34 AM

1.fork 最新代码	切换分支流程
2.修改代码	1.git branch 查看分支
3.修改ci workflow流程	2.git checkout xx 切换分支
新建workflow_tests.yml	
修改 orders_on_pr/orders_on_merge	
4.git checkout -b "add-ut-test"	
git add .	Commit 回退
Git commit -m "ssdada"	本地 git log
git config --global push.default simple	Git reset --hard xxx
git push	
git push --set-upstream origin "add-ut-test"	远程pr
	Git push -f
提交pr草稿, 多次commit, 找leader review prove	

Git pull 强制拉取并覆盖本地代码



Git 之 push 代码后, 如何回退 回退到之前的版本的方法简单整理 git push后怎么回退 -CSN博客

使用分支——Git Merge命令 - 知乎 (zhihu.com)

Git rebase
git rebase 详解 (图解+最简单易学, 一次就懂) -CSN博客

详解git rebase, 让你走上git大神之路 - 知乎 (zhihu.com)

Git——如何在 Git 中撤销 rebase? - 知乎 (zhihu.com)
git rebase, 看这一篇就够了 - 掘金 (juejin.cn)
12. Git rebase命令 git 能过某次提交的代码 -CSN博客
git rebase -i 合并多次提交 rebase 最近两次提交到另外一个分支 -CSN博客
【笔记】Git 1 将git仓库中所有的 commit 合成一个, 清空所有 git 提交记录 git 合并commit -CSN博客

git: 实现从一个仓库的指定分支合并代码到另一个仓库的指定分支 一个仓库想要切换到另一个仓库同名分支怎么办 -CSN博客

git 关于本地分支和远程分支的创建、切换、合并和删除 git 如何在本地新建分支修复后合并到本地分支 -CSN博客

git拉取远程分支到本地(两种方法)



1、本地有其他分支的代码仓库时

通过下述命令查看所有的远程分支:

```
git branch -r
```

下面有2种方法来拉取远程分支代码:

(1).需要本地分支和远程分支建立映射关系

执行如下命令:

```
git checkout -b 本地分支名xxx origin/远程分支名xxx
```

使用这种方式会在本地仓库新建分支xxx, 并自动切换到新建的分支xxx, 当前远程分支xxx的代码也拉取到了本地分支xxx中。采用这种方法建立的本地分支和远程分支建立映射关系。

(2).不需要本地分支和远程分支建立映射关系

执行如下命令:

```
git fetch origin 远程分支名xxx:本地分支名xxx
```

使用这种方式会在本地仓库新建分支xxx, 但是并不会自动切换到新建的分支xxx, 需要手动checkout, 当前远程分支xxx的代码也拉取到了本地分支xxx中。采用这种方法建立的本地分支和远程分支建立映射关系。



用两行命令删除分支

```
// 删除本地分支  
git branch -d localBranchName  
  
// 删除远程分支  
git push origin --delete remoteBranchName
```

(译者注: 关于 git push 的更多介绍, 请阅读《git push 命令的用法》)。

Basic pytest

Tuesday, January 9, 2024 4:28 PM

Pr #31

[\[workflow\] Create tests folder & workflow yml. Add tests to pr & merge orders workflow. by yutianchen666](#) • [Pull Request #31](#) • [intel/llm-on-ray \(github.com\)](#)

Pr #42

[\[Tests\] Add inference predictor tests after tests folders build by yutianchen666](#) • [Pull Request #42](#) • [intel/llm-on-ray \(github.com\)](#)

done

Pytest----如何使用--capture参数设置实时打印及Capture Log打印



DevOps技术社区 | 2022-08-28 11:28:03

目录

- 一、设置标准输出标准错误输出的模式
 - 1.1 pytest命令
 - 1.2 pytest -s 命令
 - 1.3 pytest --capture=sys 命令
 - 1.4 pytest --capture=fd 命令
 - 1.5 pytest --capture=tee-sys 命令
- 二、在测试函数中捕获标准输出标准错误输出

一、设置标准输出标准错误输出的模式

pytest捕获标准输出标准错误输出的模式主要有以下几种

- pytest：和pytest --capture=fd 模式是一样的，默认的就是pytest --capture=fd 模式

0119meeting

Friday, January 19, 2024 10:44 AM

Openai API

UT Tests之外

Tests case cover 主要的场景

子系统、不同功能自动化

测试过程中需要改进的 issue提出来 印度团队已经发现很多error了

Hf-token 设置是否有用，从config读取不成功，docker 配置映射

0119 1:1

Friday, January 19, 2024 1:04 PM

Intel 不同的方案
针对intel自己的卡工作
ray比较有机会

算法层面，考虑finetune里面的算法
Llm-on-ray里面能不能跑自己的算法和数据
数据的format不一样，是否支持
现在支持lora微调
自定义微调和层数

0122Test plan

Monday, January 22, 2024 2:02 PM

1.多个engine

hg, ipex, deepspeed, vllm支持的模型列表不相同

engine不可以叠加, ipex和vllm不兼容, deepspeed和ipex兼容, deepspeed和ipex兼容

模型支持矩阵

文档会说支持的模型, 部分

优先考虑cpu+inference

2.覆盖测试, 覆盖所有的情况

测试对应engine文档的bug, 支持矩阵

3.Feature cover

从inference开始, 第一需求

Serve.py的测试, yaml传输 case

Cover inference model

用bash脚本也可以写test

客户用命令行, 模拟命令行测试, 用pytest驱动测试, python模拟bash

4.windows docker测试

build-image.sh

测试从build docker开始, 是否生成正确image, install deps是否正确

5.test case把readme走一遍

模拟用户的install 过程, 空的conda环境

setup文档, deps是否正确安装

docker和conda分开测试

6.serve.py脚本测试

每一个参数有一个case cover

比如simple和openai是两种

Config_file, model-id serve单个模型或多个模型, 有的模型指定configfile有的支持名字即可

prefix: 非default 测试

参数可以组合case

7.query_xx.py测试

同样参数测试

新机器流程（备份）

Monday, January 22, 2024 2:52 PM

Proxy

驱动

Oneccl

Gcc -v

oneccl动态库问题 GIBxx

Dt

1source again

conda

1.登录root账户之前清除known_host信息

2.查看[Linux查看系统基本信息，版本信息（最全版）](#) [如何查看linux系统信息-CSDN博客](#)

3.新建用户和权限[linux创建新用户](#) [linux创建账户-CSDN博客](#)

4.新用户登录在 [Linux 中更改用户默认 Shell 的 3 种方法](#) ([linux-console.net](#))

5.网络设置，root账户安装net-tools apt install net-tools

5.proxy设置+dt+git设置 git、conda、apt、pip没问题 wget和ping出错

6.conda 安装[Linux下普通用户anaconda安装与配置](#) [linux的普通用户能安装conda吗-CSDN博客](#)

7.硬盘挂载 fdisk -l 查看device名称

Mkdir mnt/DP_disk1

Mount /dev/nvme0n1p1 /mnt/DP_disk1

Openai sdk问题

Thursday, January 25, 2024 6:26 PM

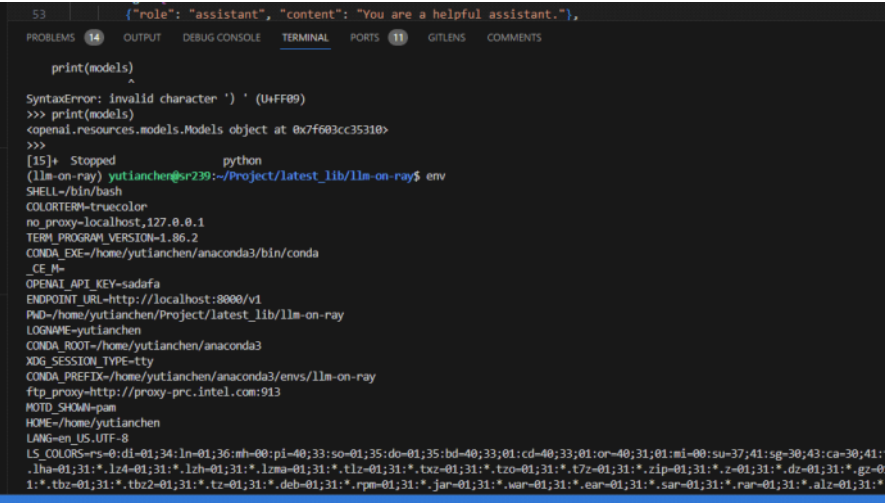
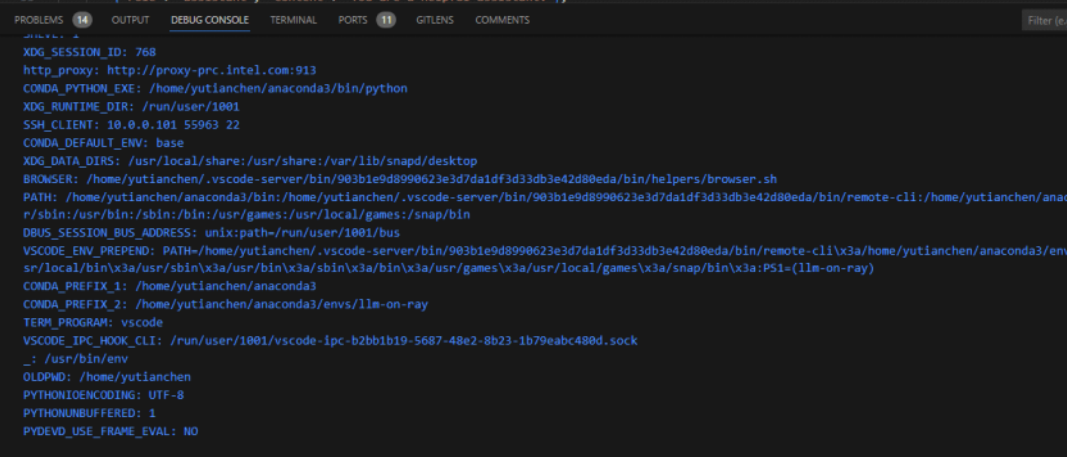
1.openai版本, [add openai version processing by KepingYan](#) • [Pull Request #80](#) • [intel/llm-on-ray](#) ([github.com](#))

2.export no_proxy=localhost,127.0.0.1
3.
"export OPENAI_API_BASE=http://localhost:8000/v1"
"export OPENAI_API_KEY=\$not_a_real_key"

环境变量

终端看得到

Debug console查不到



openai是bash里面的临时变量

需要写入整体

一、Linux的变量种类

按变量的生存周期来划分，Linux变量可分为两类：

1. 永久的：需要修改配置文件，变量永久生效。
2. 临时的：使用export命令行声明即可，变量在关闭shell时失效。

二、设置变量的三种方法

需要

Yan, Keping 1/25 6:23 PM

KY

```
export no_proxy=localhost,127.0.0.1
```

这个配置一下试试

```
PROBLEMS 14 OUTPUT DEBUG CONSOLE TERMINAL PORTS 11 GIT LENS COMMENTS 2
Find

for chunk in client.chat.completions.create(
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_utils/_utils.py", line 275, in wrapper
return func(*args, **kwargs)
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/resources/chat/completions.py", line 663, in create
return self._post(
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 1200, in post
return cast(ResponseT, self.request(cast_to, opts, stream=stream, stream_cls=stream_cls))
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 889, in request
return self._request(
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 965, in _request
return self._retry_request(
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 1013, in _retry_request
return self._request(
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 965, in _request
return self._retry_request(
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 1013, in _retry_request
return self._request(
File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 980, in _request
raise self._make_status_error_from_response(err.response) from None
openai.InternalServerError: <Idotype html>
<html>
<head>
<meta http-equiv="refresh" content="0;url=https://7rx90003.ibosscloud.com/ibereports/ibp/bp.html?bu=http://localhost:8000/v1/chat/completions&bc=Failed+to+retrieve+requested+URL.&ip=10.239.44.250&er=ERR_CONNECT_FAIL&user=328Port+916329&fp=6&cn="/>
</head>
<body>
</body>
</html>
* (llm-on-ray) (base) yutianchen@sr239:~/Project/latest_lib/llm-on-ray$ python examples/inference/api_server_openai/query_openai_sdk.py --streaming_response True
usage: query_openai_sdk.py [-h] [--model_name MODEL_NAME] [--streaming_response] [--max_new_tokens MAX_NEW_TOKENS] [--temperature TEMPERATURE] [--top_p TOP_P]
query_openai_sdk.py: error: unrecognized arguments: True
* (llm-on-ray) (base) yutianchen@sr239:~/Project/latest_lib/llm-on-ray$ python examples/inference/api_server_openai/query_openai_sdk.py -h
ChendiXue, last week In 94 Col 12 Spaces 4 UTF-8 LF
```

workflow问题-env

Tuesday, January 30, 2024 3:49 PM

```
- name: Run query Test
  run: |
    TARGET=${{steps.target.outputs.target}}
    docker exec "${TARGET}" bash -c "pip install pytest"
    docker exec "${TARGET}" bash -c "pip install openai==0.28.0"
    docker exec "${TARGET}" bash -c "export no_proxy=localhost,127.0.0.1"
    docker exec "${TARGET}" bash -c "export OPENAI_API_BASE=http://localhost:8000/v1"
    docker exec "${TARGET}" bash -c "export OPENAI_API_KEY=$not_a_real_key"
    docker exec "${TARGET}" bash -c "env"
    docker exec "${TARGET}" bash -c "./tests/run-tests.sh"
```

每一行bash分离，退出之后就没有了

```
91 container:
92   image: ${ inputs.runner_container_image }
93   env:
94     http_proxy: ${ inputs.http_proxy }
95     https_proxy: ${ inputs.https_proxy }
96     no_proxy: ${ inputs.no_proxy }
97     OPENAI_API_BASE: http://localhost:8000/v1
98     OPENAI_API_KEY: ${ inputs.OPENAI_API_KEY }
99   volumes:
100     - /var/run/docker.sock:/var/run/docker.sock
101
```

这是container的环境，用来下载container的

```
125 docker image prune -f
126
127 - name: Start Docker Container
128   run: |
129     TARGET=${{steps.target.outputs.target}}
130     cid=$(docker ps -q --filter "name=${TARGET}")
131     if [[ ! -z "$cid" ]]; then docker stop $cid && docker rm $cid; fi
132     # check and remove exited container
133     cid=$(docker ps -a -q --filter "name=${TARGET}")
134     if [[ ! -z "$cid" ]]; then docker rm $cid; fi
135     docker run -tid -v ${ inputs.model_cache_path }:/root/.cache/huggingface/hub -v ${ inputs.code_checkout_path }:/root/llm-on-ray -e http_proxy=${ inputs.http_proxy } -e https_proxy=${ inputs.https_proxy }
136
137 - name: Start Ray Cluster
138   run: |
```

要在start docker的时候添加

Checkout 把仓库代码克隆到ci机器
要看具体的checkout步骤的代码地址
运行docker时进行挂载


```
Checkout
1 ▶ Run actions/checkout@v2
12 Syncing repository: intel/llm-on-ray
13 ▶ Getting Git version info
14 Temporarily overriding HOME='/home/runner/work/_temp/31e61b79-6bb3-414c-91b6-9266fd3bea24' before making global git config changes
15 Adding repository directory to the temporary git global config as a safe directory
16 /usr/bin/git config --global --add safe.directory /home/runner/work/llm-on-ray/llm-on-ray
17 Deleting the contents of '/home/runner/work/llm-on-ray/llm-on-ray'
18 ▶ Initializing the repository
19 ▶ Disabling automatic garbage collection
20 ▶ Setting up auth
21 ▶ Fetching the repository
22 ▶ Determining the checkout info
23 ▶ Checking out the ref
24 /usr/bin/git log -1 --format='%H'
25 'db12e539e2e5f6d8bb638a4e490c6b293733131c'
```

```
13 type: string
14 default: 'http://localhost:8000/v1'
15 OPENAI_API_KEY:
16   type: string
17   default: '$your_openai_api_key'
18 code_checkout_path:
19   type: string
20   default: '/home/runner/work/llm-on-ray/llm-on-ray'
21
22 ~ jobs:
23   ~ build:
24
25     name: bare-test
```


更新源代码后import site-packages未更新问题

Tuesday, February 20, 2024 10:02 AM

然后，在包的根目录中运行以下命令：

 Copy code

```
pip install .
```

这将安装你的包及其依赖项。要更新包，只需修改代码，然后再次运行``pip install .``命令即可。

Openai1.12.0 error

Friday, February 23, 2024 2:49 PM

python examples/inference/api_server_openai/query_openai_sdk.py

```
The above exception was the direct cause of the following exception:

Traceback (most recent call last):
  File "/home/yutianchen/Project/pr_lib/llm-on-ray/examples/inference/api_server_openai/query_openai_sdk.py", line 46, in <module>
    models = client.models.list()
  File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/resources/models.py", line 79, in list
    return self.get_api_list()
  File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 1216, in get_api_list
    return self._request_api_list(model, page, opts)
  File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 1061, in _request_api_list
    return self.request(page, options, stream=False)
  File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 856, in request
    return self._request(
  File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 909, in _request
    return self._retry_request(
  File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 980, in _retry_request
    return self._request(
  File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 980, in _retry_request
    return self._request(
  File "/home/yutianchen/anaconda3/envs/llm-on-ray/lib/python3.9/site-packages/openai/_base_client.py", line 919, in _request
    raise APIConnectionError(request=request) from err
openai.APIConnectionError: Connection error.
(11m-on-ray) (base) yutianchen@sr239:~/Project/pr_lib/llm-on-ray$
```

Hg

env

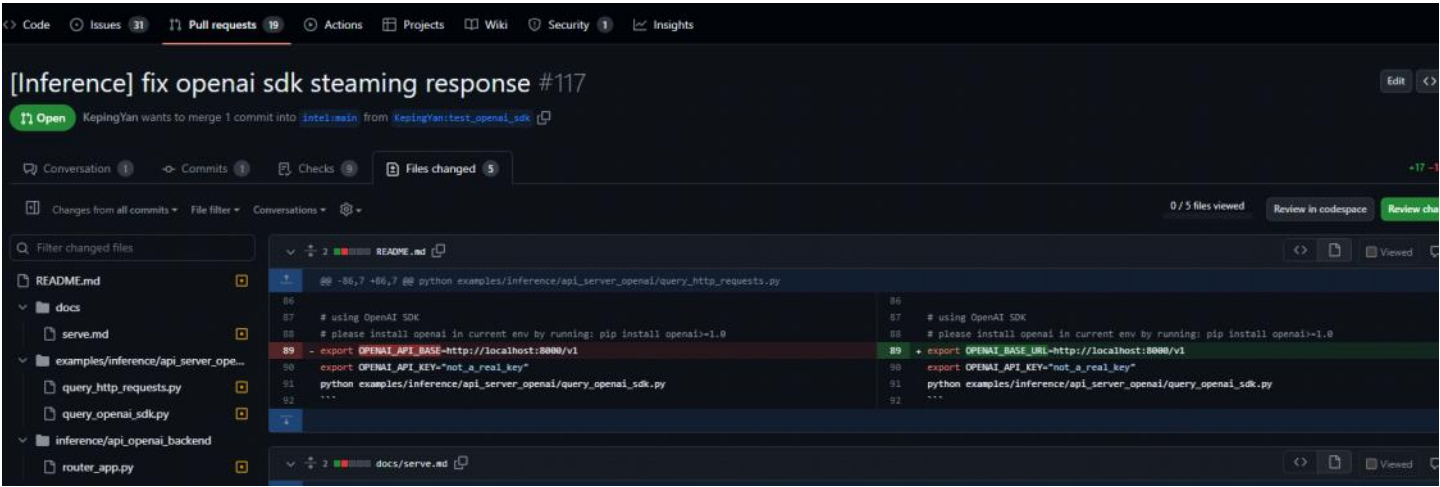
```
openai.APIConnectionError: Connection error.
(11m-on-ray) (base) yutianchen@sr239:~/Project/pr_lib/llm-on-ray$ env
SHELL=/bin/bash
COLORTERM=truecolor
no_proxy=localhost,127.0.0.1
TERM_PROGRAM_VERSION=1.86.2
CMD0_EXE=/home/yutianchen/anaconda3/bin/conda
```

```
_CE_M=
OPENAI_API_KEY=not_a_real_key
PWD=/home/yutianchen/Project/pr_lib/llm-on-
```

```
PS1=❯❯❯(11m-on-
SSH_CLIENT=10.0.
Follow link (ctrl + click)
OPENAI_API_BASE=http://localhost:8000/v1
CONDA_DEFAULT_ENV=llm-on-ray
XDG_DATA_DIRS=/usr/local/share:/usr/share:/run/user/1000/share
```

Openai=0.28.0没问题

[\[Inference\] fix openai sdk steaming response by KepingYan · Pull Request #117 · intel/llm-on-ray \(github.com\)](#)



0301meeting

Friday, March 1, 2024 10:47 AM

Table

File

File

Spreadsheet

Screen

Pictures

Online

Link

Transcribe

Record

Date

Time

Date &

Page

Equation

Symbol

Stickers

Math

Printout

Attachment

Clipping

Video

Video

Time

Time

Templates

Symbol

More

Assistant

Tables

Files

Images

Media

Links

Recording

Time Stamp

Pages

Symbols

Math

arkA21 Notebook

Events

Plan>Status.Update

Performance

ML.Key.Results

Xiaochang

review

hf --dataset ./dataset/ShareGPT_V3_unfiltered_cleaned_split.json --num-prompts 1000

Total time: 1102.20 s

Prompt Length (Min/Med/Max): 4 / 95 / 1024

Throughput: 0.91 requests/s, 433.88 tokens/s

Average latency: 556.52 s

Average latency per token: 1.97 s

current script

numactl -N 1 -m 1 python benchmark/benchmark_serving.py --model_endpoint_base http://10.1.11.2:8000/llama-2-7b-chat-hf --model_name llama-2-7b-chat-hf --dataset ./dataset/ShareGPT_V3_unfiltered_cleaned_split.json --num-prompts 1000 --dataset-format ShareGPT --results-dir ./benchmark/llmonray_throughput_2/

Total time: 618.973 s

Prompt Length (Min/Med/Max): 4 / 95 / 1024

Request Throughput (QPS): 1.616 requests/s

Token Throughput: 436.358 tokens/s

Average latency per Request: 379.226 s

Average latency per Token: 3.743 s

	32 tokens	128 tokens	512 tokens
1	<div>Total time: 1008.357 s</div> <div>Prompt Length (Min/Med/Max): 34 / 34 / 34</div> <div>Request Throughput (QPS): 0.992 requests/s</div> <div>Token Throughput: 67.436 tokens/s</div> <div>Average latency per Request: 6.117 s</div>	<div>Total time: 1040.266 s</div> <div>Prompt Length (Min/Med/Max): 137 / 137 / 137</div> <div>Request Throughput (QPS): 0.961 requests/s</div> <div>Token Throughput: 263.394 tokens/s</div> <div>Average latency per Request: 97.611 s</div>	<div>Total time: 6002.523 s</div> <div>Prompt Length (Min/Med/Max): 551 / 551 / 551</div> <div>Request Throughput (QPS): 0.167 requests/s</div> <div>Token Throughput: 183.589 tokens/s</div> <div>Average latency per Request: 2648.885 s</div>

Search

10:47 A 3/1/20

Performance test

Finetune 新文档

bash传递参数问题

Friday, March 1, 2024 4:25 PM

天辰

You

unknown flag: --a CACHEBUST\

< 2 / 2 >

ChatGPT

如果命令行解释器报告未知的标志`--a`，那可能是因为该命令不支持`--a`标志，或者`--a`不是该命令的正确用法。在这种情况下，可能需要调整命令行参数的格式。如果您想传递`CACHEBUST=1`作为参数，而不是作为标志，您可以将其作为单个参数传递，如下所示：

bash

Copy code

args+=("--a=CACHEBUST=1")

这将`CACHEBUST=1`作为`--a`的参数传递给命令。

Message ChatGPT...

+ □ □ □ ↶

第一个=作为分隔符

docker_args+=("--build-arg=CACHEBUST=1")
等价于 docker build --build-arg CACHEBUST=1

```
cp .github/workflows/scripts/docker.sh .

- name: Build Docker Image
  run: |
    DF_SUFFIX=".cpu_and_deepspeed_test"
    TARGET=${{steps.target.outputs.target}}
    PYTHON_V=${{matrix.python-version}}
    source docker.sh
    # build_and_prune TARGET=${TARGET} DF_SUFFIX=${DF_SUFFIX} PYTHON_V=${python_v} ## need change
    build_and_prune TARGET=${TARGET} DF_SUFFIX=${DF_SUFFIX} PYTHON_V=${PYTHON_V}
    # docker build ./ --build-arg CACHEBUST=1 --build-arg python_v=${{matrix.python-version}} -f dev/dockerfile
    # docker image prune -f
    docker images

- name: Start Docker Container
  run: |
```

传递参数名称相同，PYTHON_V可以，python_v不行

```
build_and_prune() {
    # Set TARGET and DF_SUFFIX using the passed in parameters
    local TARGET=$TARGET
    local DF_SUFFIX=$DF_SUFFIX
    local PYTHON_V=$PYTHON_V ## same name
    local USE_PROXY=$USE_PROXY

    echo "${PYTHON_V}"
    docker_args=()
    docker_args+=( "--build-arg=CACHEBUST=1" )
    # docker_args+=( "--build-arg=python_v=${PYTHON_V}" )
    # docker_args+=( "--build-arg=http_proxy=${HTTP_PROXY}" )
    # docker_args+=( "--build-arg=https_proxy=${HTTPS_PROXY}" )
    if [ -n "$PYTHON_V" ]; then
        docker_args+=( "--build-arg=python_v=${PYTHON_V}" )
    fi

    if [ -n "$USE_PROXY" ]; then
        docker_args+=( "--build-arg=http_proxy=${HTTP_PROXY}" )
        docker_args+=( "--build-arg=https_proxy=${HTTPS_PROXY}" )
    fi

    echo "docker build ./ ${docker_args[@]} -f dev/docker/Dockerfile${DF_SUFFIX}"

    # Build Docker image and perform cleaning operation
    docker build ./ "${docker_args[@]}" -f dev/docker/Dockerfile${DF_SUFFIX}
    docker image prune -f

    echo "docker build ./ ${docker_args[@]} -f dev/docker/Dockerfile${DF_SUFFIX}"
}
```

定义map 取消if else

```
- name: Run Inference Test with DeepSpeed
run: |
  TARGET=${{steps.target.outputs.target}}
  if [[ "${{matrix.model}}" =~ ^(gpt2|falcon-7b|starcoder|mp-7b.*)$ ]]; then
    echo "${{matrix.model}} is not supported!"
  elif [[ "${{matrix.model}}" == "llama-2-7b-chat-hf-vllm" ]]; then
    docker exec "${TARGET}" bash -c "python .github/workflows/config/update_inference_config.py --config_file llm_on_ray/
inference/models/${{matrix.model}}.yaml --output_file \"${{matrix.model}}.yaml.deepspeed --deepspeed"
    docker exec "${TARGET}" bash -c "llm_on_ray-serve --config_file \"${{matrix.model}}.yaml.deepspeed --simple"
    docker exec "${TARGET}" bash -c "python examples/inference/api_server_simple/query_single.py --model_endpoint http://
127.0.0.1:8000/${{matrix.model}})"
    docker exec "${TARGET}" bash -c "python examples/inference/api_server_simple/query_single.py --model_endpoint http://
127.0.0.1:8000/${{matrix.model}} --streaming_response"
  fi
```

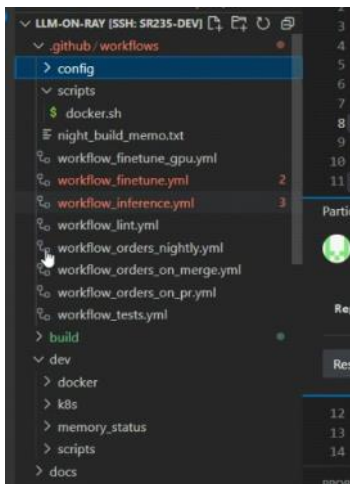
定义标准command line 如果有deepspeed, 增加--deepspeed append进去还有

类似于pytest 不同的parms, enable就append进去

定义一些开关, deepspeed,

```
source docker.sh
build_and_prune TARGET=${TARGET} DF_SUFFIX=${DF_SUFFIX} USE_PROXY=${True}
```

找一下这种传参的文档



脚本放在dev中

```
130 TARGET=${{steps.target.outputs.target}}
131 CMD=$(cat << EOF
132 import yaml
133 if "${{matrix.model}}" == "starcoder":
134     conf_path = "llm_on_ray/inference/models/starcoder.yaml"
135     with open(conf_path, encoding="utf-8") as reader:
136         result = yaml.load(reader, Loader=yaml.FullLoader)
137         result['model_description']['config']['use_auth_token'] = "${{env.HF_ACCESS_TOKEN}}"
138     with open(conf_path, 'w') as output:
139         yaml.dump(result, output, sort_keys=False)
140 if "${{matrix.model}}" == "llama-2-7b-chat-hf":
141     conf_path = "llm_on_ray/inference/models/llama-2-7b-chat-hf.yaml"
142     with open(conf_path, encoding="utf-8") as reader:
143         result = yaml.load(reader, Loader=yaml.FullLoader)
144         result['model_description']['config']['use_auth_token'] = "${{env.HF_ACCESS_TOKEN}}"
145     with open(conf_path, 'w') as output:
146         yaml.dump(result, output, sort_keys=False)
147 EOF
148 )
149 docker exec "${TARGET}" python -c "$CMD"
```

没必要写成code, 写一个文件replace template

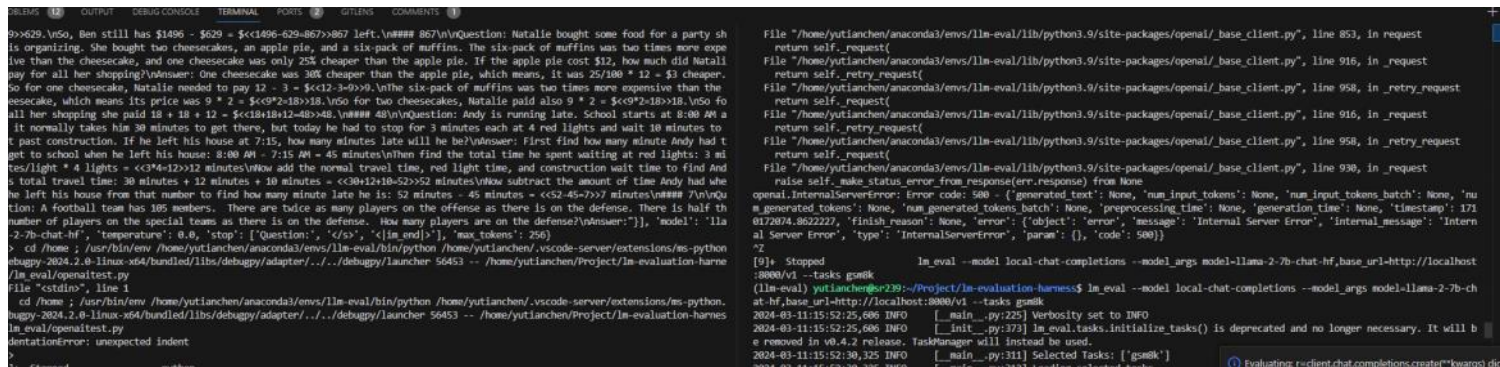
sed替换

flows/workflow_inference.yml

```
+   - name: Source build script
+     run: |
+       cp .github/workflows/scripts/docker.sh .
+
```

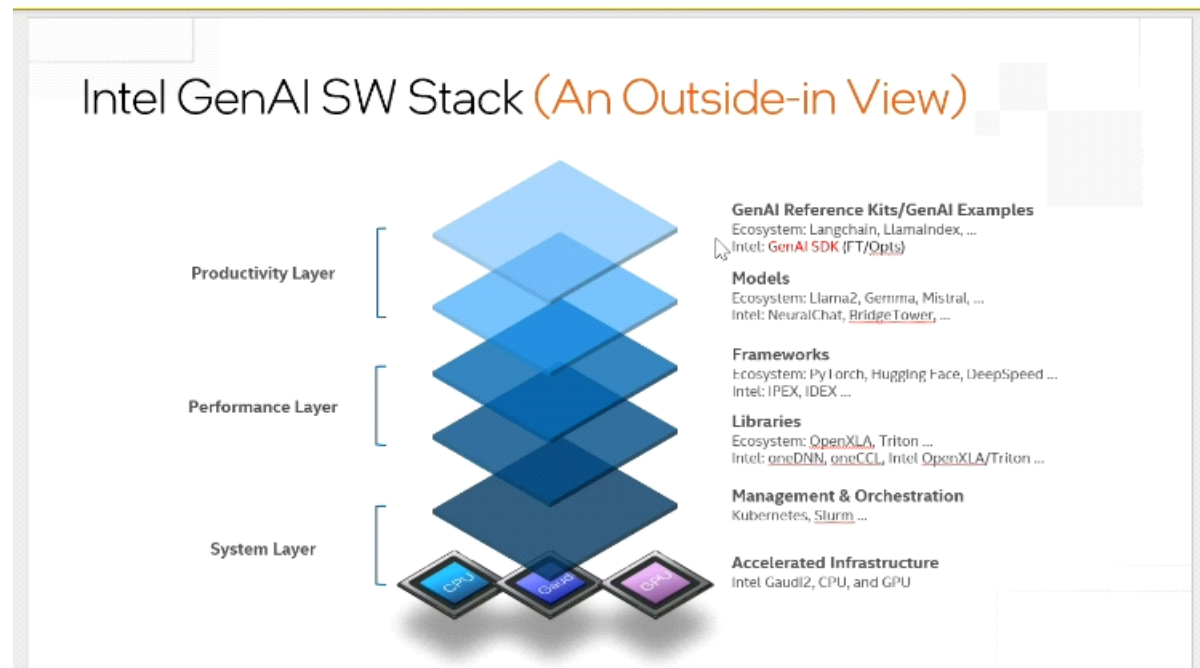
看一下是否可以直接source

Monday, March 11, 2024 4:41 PM



0315meeting

Friday, March 15, 2024 10:11 AM



[intel-sandbox/GenerativeAIEExamples](https://github.com/intel-sandbox/GenerativeAIEExamples): Generative AI Examples is a collection of GenAI examples such as ChatQnA, Copilot, which heavily reused ecosystem-based components plus Intel optimizations. (github.com)

Transformers KV cache [onenote](#).

Bigdl 是对transformers quant的封装

Itrex 底层c++没有python接口? --neuralspeed--llama.cpp

vllm、itrex、bigdl、llama.cpp

Error List Bloom-560m has error with ipex on github ci

Tuesday, March 19, 2024 1:50 PM

Github ci Cpu error avx not support for bloom-560m

0322 meeting

Friday, March 22, 2024 10:17 AM

Thursday, March 21, 2024 11:13 PM

1. output: [{"text":prompt,"input_length":365,"generate_length":4}]

llm on ray with vllm simple results:

Total time: 1074.359 s	Total time: 1129.153 s
Prompt Length (Min/Med/Max): 4 / 95 / 1024	Prompt Length (Min/Med/Max): 4 / 95 / 1024
Request Throughput (QPS): 0.931 requests/s	Request Throughput (QPS): 0.907 requests/s
Input Token Throughput: 216.961 tokens/s	Input Token Throughput: 214.192 tokens/s
output Token Throughput: 195.737 tokens/s	output Token Throughput: 142.867 tokens/s
Average latency per Request: 556.224 s	Average latency per Request: 543.929 s
Average latency per Token: 6.805 s	Average latency per Token: 42.553 s

2. data sampling:

```
sampled_indices = random_sample(range(len(dataset)),  
                                int(num_requests * 1.2))  
  
# filter  
sampled_requests = random_sample(filtered_dataset, num_requests)
```

vllm results:

Successful requests: 1000	Successful requests: 1000
Benchmark duration: 1100.526015 s	Benchmark duration: 1112.072155 s
Total input tokens: 248339	Total input tokens: 233094
Total generated tokens: 241530	Total generated tokens: 246007

throughput

Google llm model 生成prompt, 和yaml文件定义的格式 transformers 4.34增加的template

SparkA21 Notebook RayML Events Plan,Status,Update Performance ML Key Results LLM Tech Updates lianli Xianchang Minning

template, as shown in the following snippet.
Let's load the model and apply the chat template to a conversation. In this example, we'll start with a single user interaction:

```
from transformers import AutoTokenizer, AutoModelForCausalLM  
import transformers  
import torch  
model_id = "google/gemma-7b-it"  
dtype = torch.bfloat16  
tokenizer = AutoTokenizer.from_pretrained(model_id)  
model = AutoModelForCausalLM.from_pretrained(  
    model_id,  
    device_map="cuda",  
    torch_dtype=dtype,  
)  
chat = [  
    {"role": "user", "content": "Write a hello world program"},  
]  
prompt = tokenizer.apply_chat_template(chat, tokenize=False, add_generation_prompt=True)  
At this point, the prompt contains the following text:  
<bos><start_of_turn>user  
Write a hello world program<end_of_turn>  
<start_of_turn>model  
As you can see, each turn is preceded by a <start_of_turn> delimiter and then the role of the entity (either user, for content supplied by the user, or model for LLM responses). Turns finish with the <end_of_turn> token.  
You can follow this format to build the prompt manually, if you need to do it without the tokenizer's chat template.  
After the prompt is ready, generation can be performed like this:  
inputs = tokenizer.encode(prompt, add_special_tokens=False, return_tensors="pt")  
outputs = model.generate(input_ids=inputs.to(model.device), max_new_tokens=150)  
print(tokenizer.decode(outputs[0]))
```

Linux管道符

Tuesday, March 26, 2024 1:56 PM

4 人赞同了该文章

linux之管道符 ‘|’ 操作详解

管道符主要用于多重命令处理，前面命令的打印结果作为后面命令的输入。简单点说就是，就像工厂的流水线一样，进行完一道工序后，继续传送给下一道工序处理...

举个栗子：对hello.sh文件进行排序去重以后找出包含"better"的行

命令为：cat hello.sh | sort | uniq | grep 'better'

查看文本

排序

去重

过滤



Docker bare openai diff

Tuesday, April 16, 2024 4:10 PM

```
23 Service is deployed successfully.
24
25 PASSED [ 20%]
26 openai/test_openai_protocol.py::test_openai[gpt2-chat] PASSED [ 40%]
27 openai/test_openai_protocol.py::test_openai[gpt2-chat_stream] ['Config', '__abstractmethods__', '__annotations__', '__class__', '__class_vars__',
 '__config__', '__custom_root_type__', '__delattr__', '__dict__', '__dir__', '__doc__', '__eq__', '__exclude_fields__', '__fields__', '__fields_set__',
 '__format__', '__ge__', '__get_validators__', '__getattr__', '__getstate__', '__gt__', '__hash__', '__include_fields__', '__init__',
 '__init_subclass__', '__iter__', '__json_encoder__', '__le__', '__lt__', '__module__', '__ne__', '__new__', '__post_root_validators__',
 '__pre_root_validators__', '__pretty__', '__private_attributes__', '__reduce__', '__reduce_ex__', '__repr__', '__repr_args__', '__repr_name__',
 '__repr_str__', '__rich_repr__', '__schema_cache__', '__setattr__', '__setstate__', '__signature__', '__sizeof__', '__slots__', '__str__',
 '__subclasshook__', '__try_update_forward_refs__', '__validators__', 'abc_impl', 'calculate_keys', 'copy_and_set_values', 'decompose_class',
 'enforce_dict_if_root', 'get_value', 'init_private_attributes', 'iter', 'construct', 'copy', 'delta', 'dict', 'finish_reason', 'from_orm',
 'index', 'json', 'logprobs', 'model_construct', 'model_dump', 'model_dump_json', 'model_fields_set', 'parse_file', 'parse_obj', 'parse_raw', 'schema',
 'schema_json', 'to_dict', 'to_json', 'update_forward_refs', 'validate']
28 FAILED [ 60%]
29 openai/test_router.py::test_chat_completions_stream[fake_generator] PASSED [ 80%]
```

Docker

]

```
23 Serve message:
24 reading from config file, /home/runner/work/llm-on-ray/llm-on-ray/tests/openai/../../.github/workflows/config/gpt2-ci.yaml
25 Deployment is ready at `http://0.0.0.0:8000/`.
26 Service is deployed successfully.
27
28 PASSED [ 20%]
29 openai/test_openai_protocol.py::test_openai[gpt2-chat] PASSED [ 40%]
30 openai/test_openai_protocol.py::test_openai[gpt2-chat_stream] ['__abstractmethods__', '__annotations__', '__class__', '__class_getitem__',
 '__class_vars__', '__copy__', '__deepcopy__', '__delattr__', '__dict__', '__dir__', '__doc__', '__eq__', '__fields__', '__fields_set__', '__format__',
 '__ge__', '__get_pydantic_core_schema__', '__get_pydantic_json_schema__', '__getattr__', '__getattribute__', '__getstate__', '__gt__', '__hash__',
 '__init__', '__init_subclass__', '__iter__', '__le__', '__lt__', '__module__', '__ne__', '__new__', '__pretty__', '__private_attributes__',
 '__pydantic_complete__', '__pydantic_custom_init__', '__pydantic_decorators__', '__pydantic_extra__', '__pydantic_fields_set__',
 '__pydantic_generic_metadata__', '__pydantic_init_subclass__', '__pydantic_parent_namespace__', '__pydantic_post_init__', '__pydantic_private__',
 '__pydantic_root_model__', '__pydantic_serializer__', '__pydantic_validator__', '__reduce__', '__reduce_ex__', '__repr__', '__repr_args__',
 '__repr_name__', '__repr_str__', '__rich_repr__', '__setattr__', '__setstate__', '__sizeof__', '__slots__', '__str__', '__subclasshook__',
 '__weakref__', 'abc_impl', 'calculate_keys', 'check_frozen', 'copy_and_set_values', 'get_value', 'iter', 'construct', 'copy', 'delta', 'dict',
 'finish_reason', 'from_orm', 'index', 'json', 'logprobs', 'model_computed_fields', 'model_config', 'model_construct', 'model_copy', 'model_dump',
 'model_dump_json', 'model_extra', 'model_fields', 'model_fields_set', 'model_json_schema', 'model_parametrized_name', 'model_post_init',
 'model_rebuild', 'model_validate', 'model_validate_json', 'model_validate_strings', 'parse_file', 'parse_obj', 'parse_raw', 'schema', 'schema_json',
 'to_dict', 'to_json', 'update_forward_refs', 'validate']
31 PASSED [ 60%]
32 openai/test_router.py::test_chat_completions_stream[fake_generator] PASSED [ 80%]
33
```

bare

Pydantic 版本问题

```
9 Collecting openai>1.0 (from -r ./tests/requirements.txt (line 2))
10 Downloading openai-1.19.0-py3-none-any.whl.metadata (21 kB)
11 Collecting pytest-asyncio (from -r ./tests/requirements.txt (line 3))
12 Downloading pytest_asyncio-0.23.6-py3-none-any.whl.metadata (3.9 kB)
13 Collecting iniconfig (from pytest->-r ./tests/requirements.txt (line 1))
14 Downloading iniconfig-2.0.0-py3-none-any.whl.metadata (2.6 kB)
15 Requirement already satisfied: packaging in /opt/conda/lib/python3.10/site-packages (from pytest->-r ./tests/requirements.txt (line 1)) (23.2)
16 Requirement already satisfied: pluggy<2.0,>=1.4 in /opt/conda/lib/python3.10/site-packages (from pytest->-r ./tests/requirements.txt (line 1)) (1.4.0)
17 Requirement already satisfied: exceptiongroup>=1.0.0rc8 in /opt/conda/lib/python3.10/site-packages (from pytest->-r ./tests/requirements.txt (line 1)) (1.2.0)
18 Collecting tomli>=1 (from pytest->-r ./tests/requirements.txt (line 1))
19 Downloading tomli-2.0.1-py3-none-any.whl.metadata (8.9 kB)
20 Requirement already satisfied: anyio<5,>=3.5.0 in /opt/conda/lib/python3.10/site-packages (from openai>1.0->-r ./tests/requirements.txt (line 2)) (4.3.0)
21 Requirement already satisfied: distro<2,>=1.7.0 in /opt/conda/lib/python3.10/site-packages (from openai>1.0->-r ./tests/requirements.txt (line 2)) (1.8.0)
22 Collecting httpx<1,>=0.23.0 (from openai>1.0->-r ./tests/requirements.txt (line 2))
23 Downloading httpx-0.27.0-py3-none-any.whl.metadata (7.2 kB)
24 Requirement already satisfied: pydantic<3,>=1.9.0 in /opt/conda/lib/python3.10/site-packages (from openai>1.0->-r ./tests/requirements.txt (line 2)) (1.10.15)
25 Requirement already satisfied: sniffio in /opt/conda/lib/python3.10/site-packages (from openai>1.0->-r ./tests/requirements.txt (line 2)) (1.3.1)
26 Requirement already satisfied: tqdm>4 in /opt/conda/lib/python3.10/site-packages (from openai>1.0->-r ./tests/requirements.txt (line 2)) (4.66.2)
27 Requirement already satisfied: typing-extensions<5,>=4.7 in /opt/conda/lib/python3.10/site-packages (from openai>1.0->-r ./tests/requirements.txt (line 2)) (4.11.0)
28 Requirement already satisfied: idna>=2.8 in /opt/conda/lib/python3.10/site-packages (from anyio<5,>=3.5.0->openai>1.0->-r ./tests/requirements.txt (line 2)) (3.6)
29 Requirement already satisfied: certifi in /opt/conda/lib/python3.10/site-packages (from httpx<1,>=0.23.0->openai>1.0->-r ./tests/requirements.txt (line 2)) (2024.2.2)
30 Collecting httpcore==1.* (from httpx<1,>=0.23.0->openai>1.0->-r ./tests/requirements.txt (line 2))
31 Downloading httpcore-1.0.5-py3-none-any.whl.metadata (20 kB)
32 Requirement already satisfied: h11<0.15,>=0.13 in /opt/conda/lib/python3.10/site-packages (from httpcore==1.*->httpx<1,>=0.23.0->openai>1.0->-r ./tests/requirements.txt (line 2)) (0.14.0)
33 Downloading pytest-8.1.1-py3-none-any.whl (337 kB)
34
```

1.10.15不支持model_fields

```

1599 #16 16.67 Collecting huggingface-hub==0.19.4 (from datasets==2.14.6->llm-on-ray==0.0.1)
1600 #16 16.68 Downloading huggingface_hub-0.22.2-py3-none-any.whl.metadata (12 kB)
1601 #16 16.82 Collecting hjson (from deepspeed==0.11.2,==0.10.2->llm-on-ray==0.0.1)
1602 #16 16.82 Downloading hjson-3.1.0-py3-none-any.whl.metadata (2.6 kB)
1603 #16 16.98 Collecting ninja (from deepspeed==0.11.2,==0.10.2->llm-on-ray==0.0.1)
1604 #16 16.98 Downloading ninja-1.11.1.1-py2.py3-none-manylinux1_x86_64.manylinux2_5_x86_64.whl.metadata (5.3 kB)
1605 #16 17.24 Collecting pydantic<2.0.0 (from deepspeed==0.11.2,==0.10.2->llm-on-ray==0.0.1)
1606 #16 17.25 Downloading pydantic-1.10.15-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (150
kB)
1607 #16 17.25 ----- 150.6/150.6 kB 45.9 MB/s eta 0:00:00
1608 #16 17.49 Collecting click==7.0 (from ray==2.10->llm-on-ray==0.0.1)
1609 #16 17.50 Downloading click-8.1.7-py3-none-any.whl.metadata (3.0 kB)
1610 #16 17.64 Collecting jsonschema (from ray==2.10->llm-on-ray==0.0.1)
1611 #16 17.65 Downloading jsonschema-4.21.1-py3-none-any.whl.metadata (7.8 kB)
1612 #16 17.81 Collecting msgpack<2.0.0, >=1.0.0 (from ray==2.10->llm-on-ray==0.0.1)
1613 #16 17.81 Downloading msgpack-1.0.8-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (9.1
kB)
1614 #16 18.08 Collecting protobuf==3.19.5, >=3.15.3 (from ray==2.10->llm-on-ray==0.0.1)

```

deepspeed限制

```

56 Attempting uninstall: pydantic
57 Found existing installation: pydantic 1.10.15
58 Uninstalling pydantic-1.10.15:
59 Successfully uninstalled pydantic-1.10.15
60 ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. T
following dependency conflicts.
61 deepspeed 0.11.1 requires pydantic<2.0.0, but you have pydantic 2.7.0 which is incompatible.
62 WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the
to use a virtual environment instead: https://pip.pypa.io/warnings/venv
63 Successfully installed annotated-types-0.6.0 httpcore-1.0.5 httpx-0.27.0 iniconfig-2.0.0 openai-1.23.1 pydan
8.1.1 pytest-asyncio-0.23.6 tomli-2.0.1

```

Pydantic 前后对比

intel-sandbox/llm-r...

Workday

Login to EPP

ChatGPT

intel/llm-on-r...

model_fields

1/3

hide

Search

The following sections provide details on the most important changes in Pydantic V2.

Changes to `pydantic.BaseModel`

Various method names have been changed; all non-deprecated `BaseModel` methods now have names matching either the format `model_*` or `__pydantic.*__`. Where possible, we have retained the deprecated methods with their old names to help ease migration, but calling them will emit `DeprecationWarning`s.

Pydantic V1	Pydantic V2
<code>__fields__</code>	<code>model_fields</code>
<code>__private_attributes__</code>	<code>__pydantic_private__</code>
<code>__validators__</code>	<code>__pydantic_validator__</code>
<code>construct()</code>	<code>model_construct()</code>
<code>copy()</code>	<code>model_copy()</code>
<code>dict()</code>	<code>model_dump()</code>
<code>json_schema()</code>	<code>model_json_schema()</code>
<code>json()</code>	<code>model_dump_json()</code>

start_serve 放在公共代码部分

docker只装cpu，去掉deepspeed的部分，解决pydantic的问题

Refactor meeting

Friday, April 19, 2024 3:15 PM

- 1.simple 调试接口, pydantic 不同的类, get started 很快展示出来
- 2.判断最小最大版本

0423 meeting

Tuesday, April 23, 2024 3:09 PM

重点放在vllm, ipex和deepspeed不重要

vllm集成ipex kernel, 用了自己的tensor并行, 按照throughput不一定有利, 主要是latency

60/2--40?

Deepspeed tensor并行, 两个socket

Guldi支持, dynamic, vllm是否支持guldi

Production

Evaluation 只能支持transformers load model

跑一下finetune的model的效果

0510

Friday, May 10, 2024 10:10 AM

Conda 环境有自己的gcc库

Ipex+deepspeed 在bs=1时latency好

Throughput vllm更好

增加主流model支持

Issues217

Friday, May 17, 2024 1:43 PM

```
Project > pr_lib > llm-on-ray > examples > inference > api_server_simple > query_single.py > ...
42     help="The value used to modulate the next token probabilities.",
43 )
44 parser.add_argument(
45     "--top_p",
46     default=None,
47     help="If set to float < 1, only the smallest set of most probable tokens \
48         with probabilities that add up to `Top p` or higher are kept for generation.",
49 )
50 parser.add_argument(
51     "--top_k",
52     default=None,
53     help="The number of highest probability vocabulary tokens to keep \
54         for top-k-filtering.",
55 )
56
57 args = parser.parse_args()
58 prompt = "Once upon a time,"
59 config: Dict[str, Union[int, float]] = {}
60 if args.max_new_tokens:
61     config["max_new_tokens"] = int(args.max_new_tokens)
62 if args.temperature:
63     config["temperature"] = float(args.temperature)
64 if args.top_p:
65     config["top_p"] = float(args.top_p)
66 if args.top_k:
67     config["top_k"] = float(args.top_k)
68
69 sample_input = {"text": prompt, "config": config, "stream": args.streaming_response} ## define class
70
71 proxies = {"http": None, "https": None}
72 outputs = requests.post(
73     args.model_endpoint,
74     proxies=proxies, # type: ignore
75     json=sample_input, KepingYan, 4 months ago * [Lint] add lint (#34) ...
76     stream=args.streaming_response,
77 )
78
79 outputs.raise_for_status()
80 if args.streaming_response:
81     for output in outputs.iter_content(chunk_size=None, decode_unicode=True):
82         print(output, end="", flush=True)
83     print()
84 else:
85     print(outputs.text, flush=True)
86
```

```
Project > pr_lib > llm-on-ray > llm_on_ray > inference > api_openai_backend > openai_protocol.py > CompletionRequest
270 class ModelResponse(ComputedPropertyMixin, BaseModelExtended):
365     def num_total_tokens(self) -> Optional[float]:
369         return None
370
371     @property
372     def num_total_tokens_batch(self) -> Optional[float]:
373         try:
374             return (self.num_input_tokens_batch or 0) + (self.num_generated_tokens_batch or 0)
375         except Exception:
376             return None
377
378     def unpack(self) -> Tuple["ModelResponse", ...]:
379         return (self,)
380
381
382 class CompletionRequest(BaseModel):
383     model: str
384     prompt: str
385     suffix: Optional[str] = None
386     temperature: Optional[float] = None
387     top_p: Optional[float] = None
388     n: int = 1
389     max_tokens: Optional[int] = 16
390     stop: Optional[List[str]] = None
391     stream: bool = False
392     echo: Optional[bool] = False
393     presence_penalty: Optional[float] = None
394     frequency_penalty: Optional[float] = None
395     logprobs: Optional[int] = None
396     logit_bias: Optional[Dict[str, float]] = None
397     user: Optional[str] = None
398
399
400 class ChatCompletionRequest(BaseModel):
401     model: str
402     messages: List[ChatMessage]
403     temperature: Optional[float] = None
404     top_p: Optional[float] = None
405     n: int = 1
406     max_tokens: Optional[int] = None
407     stop: Optional[List[str]] = None
408     stream: bool = False
409     presence_penalty: Optional[float] = None
410     frequency_penalty: Optional[float] = None
411     logprobs: Optional[int] = None
412     logit_bias: Optional[Dict[str, float]] = None
413     user: Optional[str] = None
414     tools: Optional[List[Tool]] = None
415     tool_choice: Union[Literal["auto", "none"], ToolChoice] = "auto"
416     ignore_eos: bool = False # used in vllm engine benchmark
417
418
419 class FinishReason(str, Enum):
```

```
135         openai_serve_run(deployments, host, "/", args.port, args)
136
137     msg = "Service is deployed successfully."
138     if args.keep_serve_terminal:
139         try:
140             input(msg)
141         except EOFError:
142             return
143     else:
144         print(msg)
145
146
147 if __name__ == "__main__":
```

-it

Std in

Conda base python install 问题

Tuesday, May 21, 2024 11:38 AM

[Truststore dependency with Python 3.9 • Issue #545 • conda-forge/miniforge \(github.com\)](#)

3.9 error

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS 2

Platform: linux-64
Collecting package metadata (repodata.json): done
Solving environment: failed

LibMambaUnsatisfiableError: Encountered problems while solving:
- package mamba-1.5.8-py310h51d5547_0 requires python_abi 3.10.* *_cp310, but none of the providers can be

Could not solve for environment specs
The following packages are incompatible
├─ mamba >=1.5.8 is installable with the potential options
│   └─ mamba 1.5.8 would require
│       ├── python_abi 3.10.* *_cp310 with the potential options
│       │   ├── python_abi 3.10, which can be installed;
│       │   └─ python_abi 3.10 would require
│       │       └─ python 3.10.* *_cpython, which can be installed;
│       └─ mamba 1.5.8 would require
│           ├── libmambapy 1.5.8 py311hf2555c7_0, which requires
│           │   ├── python_abi 3.11.* *_cp311 with the potential options
│           │   │   ├── python_abi 3.11, which can be installed;
│           │   │   └─ python_abi 3.11 would require
│           │   │       └─ python 3.11.* *_cpython, which can be installed;
│           └─ mamba 1.5.8 would require
│               ├── python_abi 3.12.* *_cp312, which requires
│               │   ├── python 3.12.* *_cpython, which can be installed;
│               └─ mamba 1.5.8 would require
│                   ├── libmambapy 1.5.8 py38h5cd715c_0, which requires
│                   │   ├── python_abi 3.8.* *_cp38 with the potential options
│                   │   │   ├── python_abi 3.8, which can be installed;
│                   │   │   └─ python_abi 3.8 would require
│                   │   │       └─ python 3.8.* *_cpython, which can be installed;
│                   └─ mamba 1.5.8 would require
```

3.10 pass

Base 环境python=3.10

[Fix Miniforge3 download link in Dockerfile by controny • Pull Request #402 • aqlaboratory/openfold \(github.com\)](#)

Fix Miniforge3 download link in Dockerfile #402

Merged

jnwei merged 1 commit into aqlaboratory:main from controny:fix/miniforge3-download-link on Feb 6

Conversation 1

Commits 1

Checks 2

Files changed 1



controny commented on Feb 6

Contributor

...

I have modified the Miniforge3 download link in the Dockerfile from the latest one to a specific version. It fixes [this issue](#).

The issue is similar to [this one](#), basically caused by an inconsistency of Python version between the base environment (now becomes 3.10 due to the update of the latest Miniforge3) and what's required in `environment.yml` (3.9). So I simply tracked [the original commit upon this link](#) and replaced it with the exact version at that time.



 Fix Miniforge3 download link in Dockerfile

77916c7

Reviewers

No reviews

Assignees

No one assigned

Labels

None yet

Projects

None yet

分区 New Section 1 的第 36 页

0524meeting

Friday, May 24, 2024 10:31 AM

1.performance ppt

The screenshot shows a Microsoft Teams interface. On the left is a sidebar with navigation icons for Activity, Chat, Team, Calendar, Call, Phone, OneDrive, Setup tool, and Apps. The main area displays a document list for the 'LLM' channel in the 'General' group. The list includes the following items:

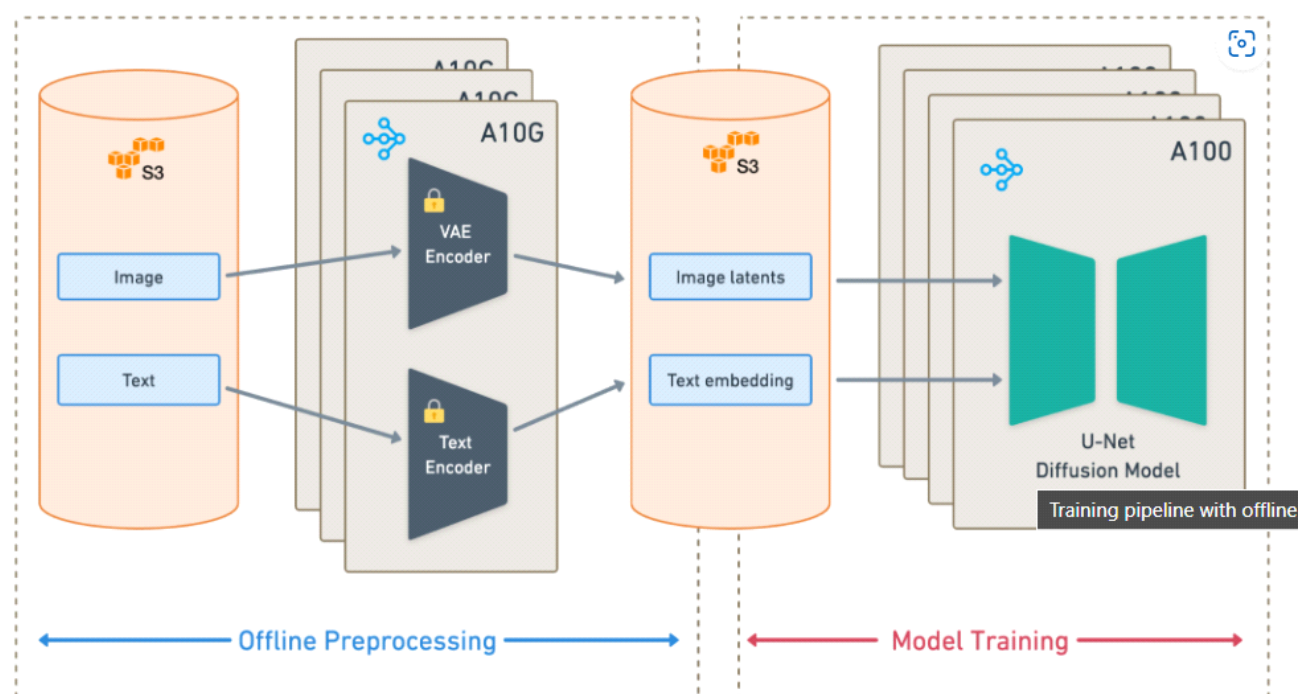
名称	修改时间	修改者
SMP-2023-ChatGLM	23年8月16日	Wang, Carson
2023_10 LLM on Ray.pptx	23年10月18日	Wang, Carson
2024_05_LLM on Ray_internal.pptx	5月13日	Yan, Keping
fine-tuning performance on HPU.pptx	5月11日	Wu, Gangsheng
LLM on Ray fine-tuning performance on PV...	3月12日	Wu, Gangsheng
LLM on Ray Roadmap.pptx	23年10月19日	Wang, Carson
The.Generative.AI.Infrastructure.Stack.jpg	23年10月10日	Wu, Xiaochang
The.Generative.AI.Market.Map.jpg	23年10月10日	Wu, Xiaochang

2.ray stable diffusion 3.7X

[Reducing the Cost of Pre-training Stable Diffusion by 3.7x with Anyscale](#)

小显卡做encoder

大显卡做unet



Streaming error

0614meeting

Friday, June 14, 2024 10:24 AM

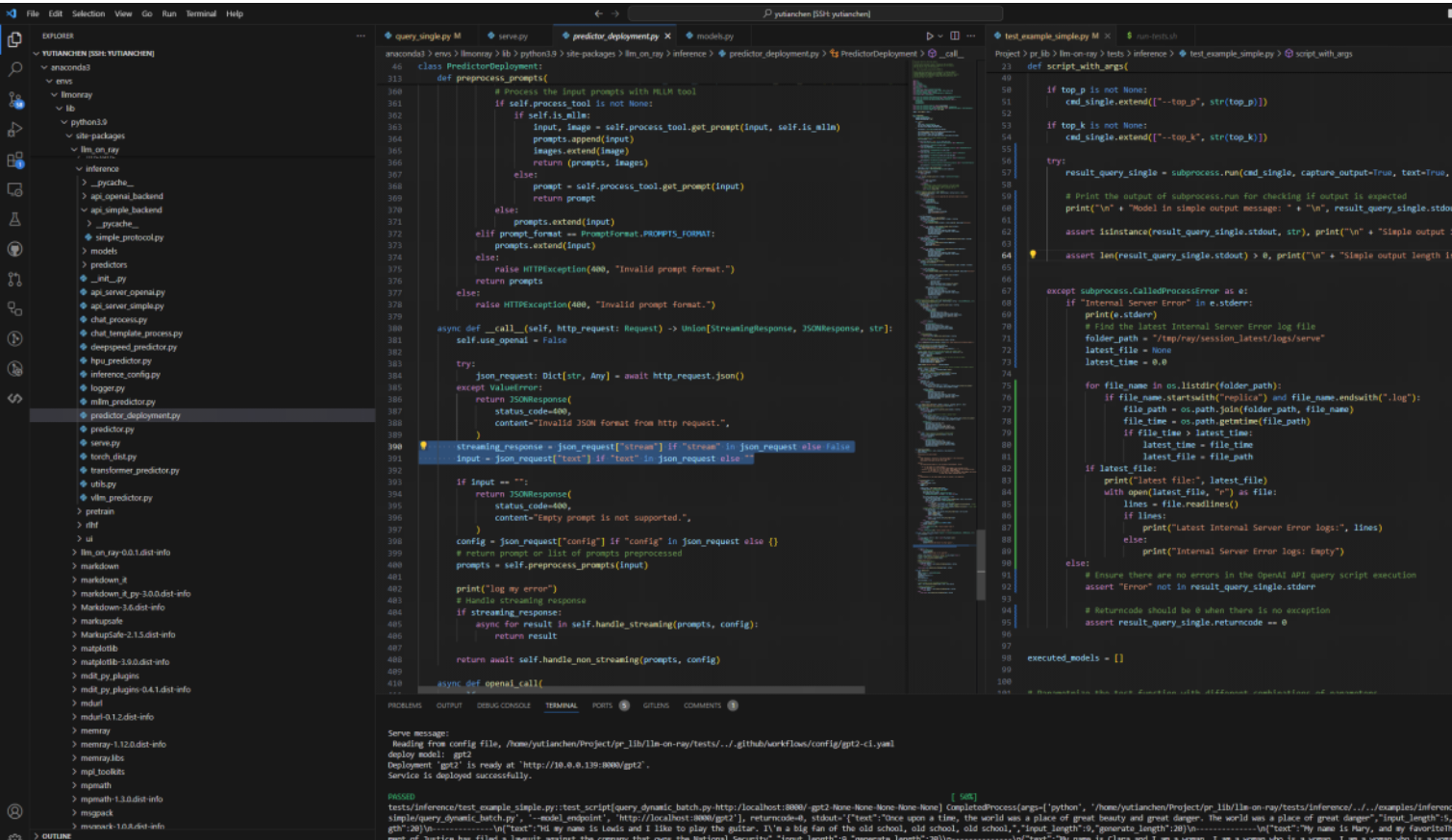
1.finetuning k8s+ray fastapi

2.vllm+ns perf

Status Update:

Pretrain: Upstream llama2 example https://github.com/ray-project/ray/pull/45459	Gangsheng	Pending on review.
Pretrain/Finetune: Support FP8 training	Gangsheng	Testing, updated habana driver, issue on memory allocation No official guide for FP8, need experiment.
Pretrain: Refactor llm-on-ray to single script	Gangsheng	Not started
Finetune: Improve evaluation result for SFT	Minming, Gangsheng	Root cause: no padding in our code and need to fill attention mask
Finetune: Support DPO and get expected evaluation result	Minming	PR submitted. Fixing CI.
Finetune: Support output Throughput metrics: Samples/s	Gangsheng	Several metrics for throughput, sentence/token per sec.
OPEA Finetuning Service	Xiaochang	
vLLM: Upstream Gaudi/CPU examples to Ray	Keping	[HPU] [Serve] Add vllm HPU support in vllm example
vLLM Gaudi + Ray benchmark	Keping	
vLLM + NS	Jiafu	Improved perf (latency and throughput) by reducing remote numa access. 8 batch: 32 users/75ms/375 tokens 4 batch: 16 users/54ms/238 tokens
Gaudi CI	Yizhong	
Refactor ci and tests	Tianchen	2PRs ready, PR244 , PR221 , one will ready today PR243 , PR219 need more review
Resnet: QA issue on Gaudi		
Upstream Vit Gaudi example https://github.com/ray-project/ray/pull/45381		
SRF, GNR Launch support for Ray		

Server 端raise一个500 错误



Pydantic 校验输入

输入转成pydantic对象

失败raise input error

参数校验 类型, optional

```
392
393
394     if input == "":
395         return JSONResponse(
396             status_code=400,
397             content="Empty prompt is not supported.",
398         )
399
```

手动校验改为pydantic

加上类型判断

增加pytest 错误

Deepspeed 需要gcc-11? ?

Friday, June 21, 2024 4:38 PM

Debug review md 格式问题

Friday, July 5, 2024 9:29 AM

- 1.bash 脚本用一个#
- 2.不要放同一行注释
- 3.代码块可以直接运行，注释放在外面
- 4.exec用脚本或者port端口映射
- 5.脚本要高亮
- 6.增加一步source docker bash

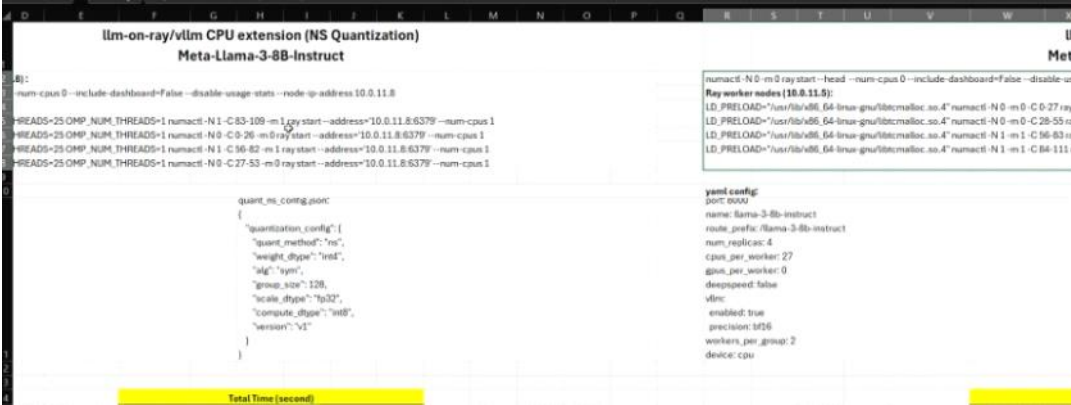
0712meeting

Friday, July 12, 2024 10:03 AM

Ns qt perfamance
Llama 3超过6144性能差
Ns 限制 超过6144之后不走flash attention cache
去掉后性能提高很多

Vllm page attention

Ray 绑定core少, 减少颠簸, 10s 500s



Vllm extension ns

Zhang, Jiatu 4:56 PM 翻译

quantization: 153 tokens 6 second

Here is a concise abstract summary of the meeting:

The meeting discussed the topic of IT security, with Chakri Krishna seeking to extend the scope of Kyle Short's previous work on security. Kyle confirmed that the focus is on IT security. The group decided to skip the one-on-one meetings with Jeevan and Rich, as Vinay Kumar had already spoken with them. Chakri Krishna asked Kyle to start from the beginning and draw a diagram to help visualize the project's scope. The meeting aimed to clarify the project's direction and build upon Kyle's previous work.

The people involved in the conversation are:

1. Kyle Short
2. Vinay Kumar
3. Chakri Krishna
4. Rich Rodgers

翻译

no-quantization: 184 tokens 18 second

Here is a concise abstract summary of the meeting:

The meeting discusses the upcoming presentation to Shredder, a new team, and the need to introduce Jeevan and Rich, who will be new faces to the team. Kyle Short confirms that IT security is the correct department to focus on. Chakri Krishna suggests creating a whiteboard diagram to help visualize the project and get everyone on the same page. Vinay Kumar suggests starting from the beginning, referencing the initial diagram, and Kyle Short notes that they spent 4 hours discussing the same diagram with security earlier. The meeting aims to clarify the scope and direction of the project.

People involved in the conversation:

不同的iteration 有些变化, 下面的是quantization的另外一个结果 194 tokens 7.6 seconds

翻译

Here is a concise abstract summary of the meeting:

The meeting involves a discussion between Kyle Short, Chakri Krishna, Vinay Kumar, Rich Rodgers, and Jason Rawlins. The main topic is to extend the work on security, with a focus on IT security. Kyle Short has already spent 4 hours with the security team and has a high-level diagram to share. Chakri Krishna wants to use this opportunity to clarify the scope and draw a whiteboard diagram to help visualize the project. Vinay Kumar has already talked to Jeevan and Rich, and Kyle Short suggests skipping the one-on-one meetings to focus on the security extension. The meeting concludes with a plan to start from the high-level diagram and work together to draw a clearer picture of the project.

The people involved in the conversation are:

1. Kyle Short
2. Chakri Krishna
3. Vinay Kumar
4. Rich Rodgers
5. Jason Rawlins

1. openvino vs. NS
2. 量化之后的效果对比
3. 理想是NS合到vllm里面

This PR is to replace the closed PR, [#264](#), which is from old branch. This PR merged some enhancements from NS main branch.

1. reshaped neural-speed as a full functional inference engine for vllm
2. integrated vllm ns extension into llm-on-ray and optimized deployment with ray
3. optimized neural-speed in several places, including compute graph construction, multiple numa node deployment and enabling flash attention kernel on llama-3-8b.
4. updated and fixed some benchmark script for IDC test and open-ai mode test, including multiple messages with different roles, removing empty chunk, fixing wrong first token latency and next token latency in open-ai mode.
5. only Llama-2-7b-chat-hf and Llama-3-8b-instruct are supported. But it can quickly extend to support other models.
6. addressed some review comments in last closed PR.
7. 2X perf improvement compared to plain vLLM cpu.

来自 <<https://github.com/intel/llm-on-ray/pull/267>>

0719 meeting

Friday, July 19, 2024 10:12 AM

4900 token/s

主要看Vllm benchmark 脚本 throughput

没有latency?

Test pr267 用户角度，如何部署

Msr-4

Friday, April 26, 2024 4:51 PM

April

Summary:

1. Update ipex-llm instead of bigdl
2. Finish better inference test case
3. Setup openai protocol and router test
4. Simplify workflows on github to bash shell

Detail:

1. PR1xx,
2. PR1xx,
3. PR111
- 4.
- 5.