

LightLLM：轻量级大型语言模型推理框架

原创

邵玫婷

于 2024-08-10 08:12:35 发布

阅读量478

收藏 3

点赞数 5

LightLLM：轻量级大型语言模型推理框架

lightllm

LightLLM is a Python-based LLM (Large Language Model) inference and serving framework, notable for its lightweight design, easy scalability, and high-speed performance.

项目地址:https://gitcode.com/gh_mirrors/li/lightllm

1. 项目介绍

LightLLM 是一款基于Python构建的大型语言模型（LLM）推理与服务框架。它的设计亮点包括轻量级架构、轻松扩展性和高性能。这个框架利用FasterTransformer、TGI vLLM和FlashAttention等优秀开源实现的优点，提供以下特色功能：

- **三进程异步协作**：令牌化、模型推断和脱标处理分别在不同进程中异步执行，提升GPU利用率。
- **Nopad（无填充）**：支持多模型的无填充注意力操作，有效处理长度差异大的请求。
- **动态批处理**：动态调度请求批次，优化资源利用率。

2. 项目快速启动

首先，确保已安装lightllm。若未安装，可使用如下命令：

```
pip install lightllm
```

然后，启动一个本地服务，假设您已有一个模型目录 /path/to/your/model：

```
1 python -m lightllm server api_server \
2     --model_dir /path/to/your/model \
3     --host 0.0.0.0 \
4     --port 1030 \
5     --nccl_port 2066 \
6     --max_req_input_len 4096 \
7     --max_req_total_len 6144 \
8     --tp 1 \
9     --trust_remote_code \
10    --max_total_token_num 120000
```

这会在本地启动一个服务，监听 1030 端口，使用 nccl_port 2066 用于NCCL通信。您可以根据实际需求调整参数。

验证服务是否正常运行，可以使用简单的HTTP请求测试：

```
1 import time
2 import requests
3 import json
4
5 url = 'http://localhost:8080/generate'
6 headers = {'Content-Type': 'application/json'}
7 payload = {
8     "inputs": "你好，世界",
9     "top_k": 5,
10    "temperature": 0.7
11 }
12
13 response = requests.post(url, json=payload, headers=headers)
14 print(response.text)
```

3. 应用案例和最佳实践

邵玫婷

关注

5

3

0

分享

打赏

...

```
1 from lazyllm import TrainableModule 2
3 # 下载并部署模型
4 m = TrainableModule('my_model')
5 deploy(m, deploy='lightllm')
6
7 # 启动服务
8 start()
9 wait()
10
11 # 关闭服务
12 stop()
```

- **效率优化**: 对于大规模模型，通过增加tp参数，利用TensorParallel在多张GPU上 **并行** 推断。
- **安全性设置**: 在生产环境中，谨慎使用--trust_remote_code选项，以防止不受信任的远程代码执行。

4. 典型生态项目

- **FasterTransformer**: 高性能的Transformer计算库。
- **TGI vLLM**: TensorFlow实现的在线微调 and 推理框架。
- **FlashAttention**: 加速Transformer中自注意力层的计算库。

以上即为LightLLM的简介及使用入门，更多详细信息和进阶教程，请参考官方文档。

lightllm

LightLLM is a Python-based LLM (Large Language Model) inference and serving framework, notable for its lightweight design, easy scalability, and high-speed performance.

项目地址:https://gitcode.com/gh_mirrors/li/lightllm

大语言模型原理基础与前沿 轻量级微调 AI天才研

1. 背景介绍 随着深度学习技术的不断发展，大语言模型 (large language models, LLM) 已经成为人工智能领域的热门研究方向之一。这些模型可以通过自监督方式学习。

LightLLM 速览 步子哥de

LightLLM是一个纯Python的超轻量高性能LLM推理框架，用于部署大语言模型并实现高吞吐量的推理服务。它解决了大模型推理部署中的一些挑战，如显存碎片化、请求排队等。

...LLM并发加速部署方案(llama.cpp、vllm、lightLLM、fastLLM)_llamac...

vllm:基于Python,①PagedAttention高效管理注意力KV内存,②连续动态批处理,③量化GPTQ/AWQ/SqueezeLLM等。 lightllm:基于Python,①三进程异步协作,②动态批处理,

什么是大模型框架?常用的大模型框架盘点对比_大模型 框架

3、LightLLM 4、llama.cpp 5、LocalAI 6、适用昇腾AI处理器的框架 7、fastllm 8、DeepSpeed-MII 9、TensorRT-LLM 10、其他。LM Studio、xinference、Colossal-AI等

模型压缩与加速：轻量级AI大语言模型的设计与实现 AI天才研

1. 背景介绍 1.1 大型AI语言模型的崛起 近年来，随着深度学习技术的快速发展，大型AI语言模型如GPT-3、BERT等在自然语言处理（NLP）领域取得了显著的成果。这些

LightLLM：轻量级、高性能的语言模型推理框架 gitblog_00585de

LightLLM：轻量级、高性能的语言模型推理框架 项目地址:https://gitcode.com/gh_mirrors/li/lightllm 在人工智能领域，大型语言模型（LLM）的应用日益广泛，但高推理延迟和显存碎片化是其主要挑战。

...大模型框架vLLM、大模型框架LightLLM、大模型框架llama.cpp、大模型...

常见的大模型框架:大模型框架Ollama、大模型框架vLLM、大模型框架LightLLM、大模型框架llama.cpp、大模型框架LocalAI、大模型框架veGiantModel: 大模型框架是指

不同的llm推理框架_taco-llm

不同的llm推理框架 vLLM适用于大批量Prompt输入,并对推理速度要求比较高的场景。 实际应用场景中,TensorRT-LLM通常与Triton InferenceServer结合起来使用,NVIDIAI

探索LightLM：一个高效轻量级的语言模型 gitblog_00004de

探索LightLM：一个高效轻量级的语言模型 去发现同类优质开源项目:<https://gitcode.com/> 是由CLUE Benchmark团队开发的一个创新性的语言模型，旨在提供高性能、低推理延迟。

LightLLM: 大型语言模型推理与服务框架入门指南 gitblog_00911de

LightLLM: 大型语言模型推理与服务框架入门指南 项目地址:https://gitcode.com/gh_mirrors/li/lightllm 目录结构及介绍 根目录 README.md: 项目的主要介绍，包括许可协议

...推理加速框架及用法(vLLM/DeepSpeed-MII/LightLLM/TensorRT-LLM)-CS...

1.1 vLLM 1.1.1 简介 vLLM是一个快速且易于使用的LLM推理和服务库。vLLM具有以下特点: (1) 先进的服务吞吐量; (2) PagedAttention对注意力Keys和Values存储的有效管理。

【LLM】Gemma：最新轻量级开源大语言模型实践 2301_81888214de

【LLM】Gemma：最新轻量级开源大语言模型实践

ART框架自动多步推理与工具利用提升大型语言模型能力 最新发布

邵玫婷 关注

5 3 0 分享 打赏

任务