PR #23047 完整报告

sgl-project/sglang

[Lora] Support LoRA and multi-batch in bench_one_batch_server

合并时间：2026-04-22 05:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23047

执行摘要

- 一句话：为 bench_one_batch_server 添加 LoRA 和多批处理支持
- 推荐动作：对于从事 LoRA 性能基准测试的工程师，该 PR 值得精读。其中关于多批处理模式的设计（跳过 max_running_requests 检查、重新计算 token_capacity）以及 LoRA 请求分布策略（uniform/distinct/skewed）是值得关注的决策。

功能与动机

现有的 bench_one_batch_server 无法对预加载的 LoRA 适配器进行基准测试，也没有方式通过多批处理摊提每次运行的开销以获得更稳定的吞吐量数据。

实现拆解

1. 在 BenchArgs 数据类中添加 lora_name、lora_request_distribution、lora_zipf_alpha 和 enable_multi_batch 字段，并注册对应的 CLI 参数。
2. 在 run_one_case 函数中增加 lora_name 等参数，根据 lora_name 和 lora_request_distribution 生成每个请求的 LoRA 路径字段，并在请求体中设置 lora_path。
3. 在 run_one_case 中实现多批处理逻辑：当 enable_multi_batch 为 True 时，跳过 max_running_requests 检查，并调整 token_capacity 检查以使用 min(batch_size, running_cap) * (input_len + output_len)，保证 OOM 保护。
4. 更新文档 docs/developer_guide/benchmark_and_profiling.md，记录 --enable-multi-batch 和 --lora-name 的使用说明。

关键文件：

- python/sglang/test/bench_one_batch_server_internal.py（模块 基准测试；类别 test；类型 test-coverage；符号 BenchArgs, add_cli_args, run_one_case）：核心变更文件，添加了 LoRA 和多批处理相关的 CLI 参数及执行逻辑。
- docs/developer_guide/benchmark_and_profiling.md（模块 文档；类别 docs；类型 documentation）：文档更新，说明新增参数的使用方法和注意事项。

关键符号：run_one_case, add_cli_args, BenchArgs

关键源码片段

python/sglang/test/bench_one_batch_server_internal.py

核心变更文件，添加了 LoRA 和多批处理相关的 CLI 参数及执行逻辑。

```
# 在 BenchArgs 数据类中新增的字段
lora_name: Optional[List[str]] = None # 指定预加载的 LoRA 适配器名称列表
lora_request_distribution: str = "uniform" # 多适配器时的分配策略：uniform / distinct / skewed
lora_zipf_alpha: float = 1.1 # skewed 分布时的 Zipf 参数
enable_multi_batch: bool = False # 启用多批处理模式

# 在 add_cli_args 中注册对应的 CLI 参数
parser.add_argument(
    "--lora-name",
    type=str,
    nargs="*",
    default=BenchArgs.lora_name,
    help="预加载的 LoRA 适配器名称，将作为 lora_path 字段发送。要求服务端已启用 LoRA。"
)
parser.add_argument(
    "--lora-request-distribution",
    type=str,
    default=BenchArgs.lora_request_distribution,
    choices=["uniform", "distinct", "skewed"],
    help="多适配器时如何为每个请求选择适配器。"
)
parser.add_argument(
    "--lora-zipf-alpha",
    type=float,
    default=BenchArgs.lora_zipf_alpha,
    help="当使用 skewed 分布时，控制 Zipf 分布的 alpha 参数。"
)
parser.add_argument(
    "--enable-multi-batch",
    action="store_true",
    help="允许 --batch-size 超过服务端 running_cap，剩余请求排队依次处理，仅 overall_
    throughput 具有参考意义。"
)
```

评论区精华

该 PR 未收到实质性的 review 讨论，仅 bot 自动评论确认没有需要反馈的内容，随后由 hnyls2002 批准合并。

- 暂无高价值评论线程

风险与影响

- 风险：由于变更仅限于 bench_one_batch_server 基准测试工具，且默认行为保持不变，对核心服务无影响。主要风险在于新参数可能与其他参数组合产生未预料的行为，但通过测试验证和文档说明可降低风险。

- 影响：用户可通过新参数更便捷地对 LoRA 适配器进行性能基准测试，并利用多批处理模式获得更稳定的吞吐量测量值。对团队而言，新增代码量小，维护成本低，但需注意 future 修改 bench_one_batch_server 时保持兼容。
- 风险标记：低影响范围，测试工具变更

关联脉络

- 暂无明显关联 PR