PR #32435 完整报告

sgl-project/sglang

Load initial expert location metadata on CPU

合并时间：2026-07-26 20:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32435

执行摘要

- 一句话：Expert location 元数据加载时指定 CPU
- 推荐动作：值得合并的小修复，逻辑清晰，风险低。维护者应确保后续代码对张量设备位置的处理正确。

功能与动机

PR body 指出 torch.load 默认将张量恢复到保存时记录的设备，对于 expert location 元数据（配置元数据），可能导致不必要的设备初始化、设备不匹配错误或在不同硬件 /rank 配置下加载失败。将数据加载到 CPU 更合理，因为它是配置元数据，无需直接加载到加速器设备。

实现拆解

1. 在 python/sglang/srt/eplb/expert_location.py 文件的 compute_initial_expert_location_metadata 函数中，当加载 .pt 文件时，将 torch.load(data, weights_only=True) 修改为 torch.load(data, weights_only=True, map_location="cpu")。
2. 该改动确保所有从 .pt 文件加载的 expert location 元数据都放在 CPU 上，而非默认的设备。

关键文件：

- python/sglang/srt/eplb/expert_location.py（模块 专家位置；类别 source；类型 core-logic；符号 compute_initial_expert_location_metadata）：核心修改文件，修复 torch.load 默认加载设备问题

关键符号：compute_initial_expert_location_metadata

关键源码片段

python/sglang/srt/eplb/expert_location.py

核心修改文件，修复 torch.load 默认加载设备问题

```
def compute_initial_expert_location_metadata(
    server_args: ServerArgs,
    model_config: ModelConfig,
    moe_ep_rank: int,
```

```
) -> Optional[ExpertLocationMetadata]:
    data = server_args.init_expert_location
    if data == "trivial":
        return ExpertLocationMetadata.init_trivial(
            server_args, model_config, moe_ep_rank
        )

    # TODO unify with the utils function
    if data.endswith(".pt"):
        # 关键修复：map_location="cpu" 确保元数据加载到 CPU，
        # 避免 torch.load 默认恢复到保存时的设备（如 GPU）
        data_dict = torch.load(data, weights_only=True, map_location="cpu")
    elif data.endswith(".json"):
        data_dict = json.loads(Path(data).read_text())
    else:
        data_dict = json.loads(data)

    if "physical_to_logical_map" in data_dict:
        logger.info(
            "init_expert_location from init_by_mapping using ServerArgs.init_expert_location"
        )
        return ExpertLocationMetadata.init_by_mapping(
            server_args,
            model_config,
            **data_dict,
            moe_ep_rank=moe_ep_rank,
        )
    elif "logical_count" in data_dict:
        logger.info(
            "init_expert_location from init_by_eplb using ServerArgs.init_expert_location"
        )
        return ExpertLocationMetadata.init_by_eplb(
            server_args, model_config, logical_count=data_dict["logical_count"]
        )
    else:
        raise NotImplementedError(
            f"Unknown init_expert_location format ({list(data_dict.keys())=})"
        )
```

评论区精华

无 review 评论。作者 sglang-npu-bot 在 issue 评论中说明仅修改了静态 eplb 函数，由于 GPU 环境故障未能运行全部测试，但分析认为不会影响其他场景。

- 暂无高价值评论线程

风险与影响

- 风险：风险很低。仅修改了一行代码，且改动明确（添加 map_location="cpu"）。可能的风险是如果调用方后续期望张量在特定设备上（例如 GPU），则可能出现设备不匹配。但根据 PR 上下文和代码逻辑，此元数据仅用于配置，后续使用会通过 .cpu() 或其他方式处理，因此风险极低。
- 影响：影响范围极小，仅影响 .pt 格式的 expert location 元数据加载路径。用户加载此类元数据时不再默认占用 GPU 内存，减少显存浪费，提高跨硬件兼容性。
- 风险标记：缺少测试覆盖

关联脉络

- PR #24256 [core/loader] Add presharded load format: 同样涉及模型加载性能优化，但具体功能不同