

PR #30323 完整报告

sgl-project/sglang

Use FP32 logits in MoEGate fallbacks

合并时间: 2026-07-08 23:43

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30323>

执行摘要

- 一句话: 确保 MoE 门控回退路径使用 FP32 logits
- 推荐动作: 值得精读以理解 SGLang 中 MoE 路由器各个精度路径的设计考量。展示了如何在尽量不引入额外性能开销的前提下修复数值精度漏洞。

功能与动机

29783 已处理了大部分路由器 GEMM 的 BF16→FP32 精度问题, 但 prefill-CP 回退分支仍使用 `F.linear`, 产生 BF16 的 router logits, 导致 NVFP4 量化时精度下降。该 PR 补全这一遗漏路径, 确保所有 CUDA 上的路由器 GEMM 输出均为 FP32。

实现拆解

1. 定位遗漏路径: 在 `DeepseekV2Gate.forward` 中, prefill-CP 分支 (针对非 DeepSeekV4 且启用 context parallelism 的情况) 原本直接调用 `F.linear`, 该函数输出 dtype 与输入一致, 若输入为 BF16 则输出 BF16, 导致精度损失。
2. 引入 FP32 内核: 为 CUDA 后端导入 `sglang.jit_kernel.dsv4.linear_bf16_fp32`, 该内核在 BF16 计算但输出 FP32, 与 #29783 中其他路径的语义一致。非 CUDA 后端 (如 CPU) 保留 `F.linear` 回退。
3. 保留确定性分支: `enable_deterministic_inference` 分支不受影响, 仍使用 `F.linear`, 因为 `linear_bf16_fp32` 虽然位精确但可能导致较敏感的 routing 变动。
4. 测试覆盖: 未新增独立测试, 依赖现有的 prefill-CP 集成测试 (如 `test_deepseek_v3_cp_single_node.py`) 验证功能。

关键文件:

- `python/sglang/srt/models/deepseek_v2.py` (模块 MoE 门控; 类别 source; 类型 data-contract): 唯一变更文件, 修改了 `DeepseekV2Gate` 的路由器 GEMM 前向逻辑, 补充 prefill-CP 回退分支的 FP32 输出。

关键符号: `DeepseekV2Gate.forward`

关键源码片段

python/sglang/srt/models/deepseek_v2.py

唯一变更文件，修改了 DeepseekV2Gate 的路由器 GEMM 前向逻辑，补充 prefill-CP 回退分支的 FP32 输出。

```
# python/sglang/srt/models/deepseek_v2.py — DeepseekV2Gate.forward
# 修改点：prefill-CP 回退分支（原仅 F.linear）现在对 CUDA 使用 linear_bf16_fp32 产生 FP32 logits

if (
    not self.is_deepseek_v4
    and forward_batch is not None
    and (
        dsa_use_prefill_cp(forward_batch, self.dsa_enable_prefill_cp)
        or mla_use_prefill_cp(forward_batch, self.mla_enable_prefill_cp)
    )
):
    if _is_cuda:
        # 导入定制 kernel: BF16 matmul + FP32 输出，保证路由器 logits 精度
        from sglang.jit_kernel.dsv4 import linear_bf16_fp32
        return linear_bf16_fp32(hidden_states, self.weight)
        # 非 CUDA 后端 (CPU/others) 仍使用标准 linear，输出 BF16
        return F.linear(hidden_states, self.weight, None)
    else:
        # 其他路径已通过 #29783 保证 FP32，此处省略细节 ...
        pass

return logits
```

评论区精华

- b8zhong确认了 FP32 bias 的决策：“Let's keep the bias in f32. Since the correction bias has sensitive values”。
- Fridge003最初要求不要在确定性分支上应用改动（“We shouldn't change here since deterministic inference relies on F.linear”），mmangkad验证后回退了该部分并解释“linear_bf16_fp32 rounds back bit-exactly to F.linear, but the FP32 logits can still change top-k routing before rounding”。最终 Fridge003 批准。
- 确定性分支不应被改动 (correctness): 确定性分支保持原样，仅修改 prefill-CP 分支。
- 保持 bias 为 FP32 的决策 (design): 后续所有路由器 GEMM 输出均为 FP32。

风险与影响

- 风险：风险极低。变更仅影响 DeepseekV2Gate.forward 中 prefill-CP 分支的一个条件内，且仅在 CUDA 后端替换为语义等价但输出 dtype 提升的 kernel。非 CUDA 路径无行为变化。唯一潜在风险是 linear_bf16_fp32 对 inputs 有对齐要求（hidden_states.shape[1] % 1024 == 0），但该分支之前的条件已隐含了类似假设。确定性分支已确保不受影响。
- 影响：

- 用户层面：使用 DeepSeekV2/V3 模型且启用 prefill-CP 的用户将获得更稳定的路由决策，尤其对 NVFP4 量化有可观测的 accept length 提升（参考 #29783 的数据）。
- 系统层面：无性能退化，linear_bf16_fp32 是定制优化 kernel，与 F.linear 相比计算量相同但输出 dtype 转换开销可忽略。
- 团队层面：厘清了所有 router GEMM 路径的 dtype 契约，降低后续精度问题的排查成本。
- 风险标记：暂无

关联脉络

- PR #29783 Fixes for NVFP4 numerical accuracy for router GEMM output and wrong correction bias cast: 该 PR 是 #29783 的补充，修复了其遗留的 prefill-CP 回退路径精度问题。