PR #21405 完整报告

sgl-project/sglang

Enable IndexCache for DeepSeek V3.2

合并时间：2026-04-05 17:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21405

执行摘要

本 PR 为 DeepSeek V3.2 模型启用了 IndexCache 优化，通过跨层复用 topk 索引减少计算开销，实现了约 6.4% 的吞吐量提升和约 5.5% 的延迟降低，同时引入了可配置的索引复用模式，但需注意潜在的精度损失。

功能与动机

动机源于 issue #21286，要求实现 IndexCache 以加速稀疏注意力计算。PR body 引用了论文《IndexCache: Accelerating Sparse Attention via Cross-Layer Index Reuse》，并提供了详细的性能基准测试，显示在 GSM8K 任务上精度保持稳定，性能显著提升。

实现拆解

主要修改了 DeepSeek 模型的核心文件：

- python/sglang/srt/models/deepseek_v2.py：添加 skip_topk 和 next_skip_topk 逻辑，支持 index_topk_freq 和 index_topk_pattern 配置。
- python/sglang/srt/models/deepseek_common/attention_forward_methods/forward_mla.py：修改 forward_absorb_prepare 和 forward_absorb_core 方法，在 skip_topk 为 True 时复用 prev_topk_indices。
- python/sglang/srt/models/deepseek_nextn.py：更新 forward 方法以传递 topk_indices。
- test/registered/8-gpu-models/test_deepseek_v32_indexcache.py：新增测试验证功能。

关键代码逻辑示例：

```
if not self.skip_topk or prev_topk_indices is None:
    topk_indices = self.indexer(...)
else:
    topk_indices = prev_topk_indices
```

评论区精华

review 中主要交锋点：

- arXiv 引用错误：gemini-code-assist 指出引用年份 2603 可能有误，需修正。
- 测试文件大小：Fridge003 建议优化测试，避免大型夜间测试。
- skip_topk 逻辑：Fridge003 讨论实现方式，最终参考官方补丁调整。

风险与影响

风险：

1. 精度损失：索引复用可能影响模型输出准确性，issue 评论中确认存在精度损失。
2. 兼容性问题：修改前向传播接口可能影响 TBO 路径或其他模型变体。
3. 测试覆盖不足：测试仅针对特定配置，可能遗漏边缘情况。

影响：

- 用户：性能提升但需权衡精度。
- 系统：新增配置选项，增加灵活性。
- 团队：需维护新逻辑，确保跨硬件兼容性。

关联脉络

与 PR #21502 关联，后者在 NPU 上实现了类似功能，显示了 IndexCache 优化的跨硬件扩展性。此外，与仓库中其他性能优化 PR（如 PR #21771）有共同主题，反映了团队对推理效率的持续关注。