PR #24459 完整报告

sgl-project/sglang

Register aten::rms_norm and aten::mm.dtype in batch invariant mode

合并时间：2026-05-06 08:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24459

执行摘要

- 一句话：在 batch invariant 模式中注册 rms_norm 和 mm.dtype
- 推荐动作：该 PR 逻辑清晰、改动精简，值得关注的是其对 batch-invariant 兼容层的扩展模式，展示了如何为更多 ATen 算子添加确定性支持。建议后续测试覆盖新增算子的确定性行为。

功能与动机

确保使用 aten::rms_norm 和 aten::mm.dtype 算子的模型在确定性模式（batch_invariant_mode）下行为确定，避免因 batch 合并导致的数值不一致。

实现拆解

1. 在 batch_invariant_ops.py 中新增 _get_or_make_ones 工具函数，用于缓存全 1 张量，避免重复创建。
2. 实现 _rms_norm_aten_compat 包装函数，兼容 aten::rms_norm 接口：处理 weight 和 eps 可选参数，在缺失时使用默认值（weight 为全 1 张量，eps 为 finfo(input.dtype).eps），并断言 normalized_shape 为最后一维，然后调用已有的 rms_norm_batch_invariant。
3. 实现 _mm_dtype_compat 包装函数，兼容 aten::mm.dtype 接口：调用 matmul_persistent 并对结果进行 to(out_dtype) 类型转换。
4. 在 enable_batch_invariant_mode 中通过 _batch_invariant_LIB.impl 注册这两个算子，与已有的 _log_softmax、mean.dim 等注册模式一致。

关键文件：

- python/sglang/srt/batch_invariant_ops/batch_invariant_ops.py（模块 确定性算子层；类别 infra；类型 infrastructure；符号 _get_or_make_ones, _rms_norm_aten_compat, _mm_dtype_compat）：核心变更文件：新增 _get_or_make_ones、_rms_norm_aten_compat、_mm_dtype_compat 函数，并在 enable_batch_invariant_mode 中注册这两个 ATen 算子。

关键符号：_get_or_make_ones, _rms_norm_aten_compat, _mm_dtype_compat

关键源码片段

python/sglang/srt/batch_invariant_ops/batch_invariant_ops.py

核心变更文件：新增 _get_or_make_ones、_rms_norm_aten_compat、_mm_dtype_compat 函数，并在 enable_batch_invariant_mode 中注册这两个 ATen 算子。

```
# 全 1 张量缓存，避免每次创建新张量
_ONES_CACHE: dict[Tuple, torch.Tensor] = {}

def _get_or_make_ones(shape, device, dtype) -> torch.Tensor:
    """获取或创建指定 shape/device/dtype 的全 1 张量（带缓存）"""
    key = (tuple(shape), device, dtype)
    t = _ONES_CACHE.get(key)
    if t is None:
        t = torch.ones(shape, device=device, dtype=dtype)
        _ONES_CACHE[key] = t
    return t


def _rms_norm_aten_compat(input, normalized_shape, weight=None, eps=None):
    """兼容 aten::rms_norm 的 batch-invariant 包装器"""
    if eps is None:
        eps = torch.finfo(input.dtype).eps
    if weight is None:
        # 当 weight 为 None 时使用全 1 张量
        weight = _get_or_make_ones(normalized_shape, input.device, input.dtype)
    # 仅支持最后一维归一化（与 rms_norm_batch_invariant 一致）
    assert tuple(normalized_shape) == (input.shape[-1],), (
        "rms_norm_batch_invariant only supports last-dim normalization "
        f"(got normalized_shape={tuple(normalized_shape)}, "
        f"input.shape={tuple(input.shape)})"
    )
    return rms_norm_batch_invariant(input, weight, eps=eps)


def _mm_dtype_compat(self, mat2, out_dtype):
    """兼容 aten::mm.dtype 的 batch-invariant 包装器：对齐后计算，再转 dtype"""
    return matmul_persistent(self.contiguous(), mat2.contiguous()).to(out_dtype)


# 在 enable_batch_invariant_mode 中注册这两个算子
_batch_invariant_LIB.impl("aten::rms_norm", _rms_norm_aten_compat, dispatch_key)
_batch_invariant_LIB.impl("aten::mm.dtype", _mm_dtype_compat, dispatch_key)
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 新增的 _rms_norm_aten_compat 仅支持最后一维归一化（normalized_shape == (input.shape[-1],)），若模型使用其他归一化形状将触发断言失败。
2. _mm_dtype_compat 对输入进行了 contiguous() 调用，可能引入额外内存开销；但这是正确性保证的常见做法。
3. _ONES_CACHE 缓存未考虑清理机制，长时间运行的模型可能累积张量缓存，但通常影响不大（缓存 key 有限）。 - 影响：影响范围较小：仅影响启用 batch_invariant_mode 的模型推理路径，新增两个 ATen 算子的注册，无 API 或配置变更。对使用 rms_norm 或 mm.dtype 的模型（如部分 DeepSeek 变体）提供确定性保证。 - 风险标记：缺少测试覆盖，核心路径变更

关联脉络

- PR #24392 add indexer-topk capture (V3.2 NSA + infra): 同为 batch-invariant 确定性功能增强，涉及模型 deterministic 行为。