PR #41569 完整报告

vllm-project/vllm

[ROCm][CI] Fix MLA prefill scale for DeepSeek GSM8K

合并时间：2026-05-05 07:32

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41569

执行摘要

- 一句话：修复 DeepSeek MLA prefill scale 计算错误
- 推荐动作：值得精读，尤其关注 get_mla_prefill_scale 中的模型版本区分逻辑（compress_ratios 属性）和 YaRN mscale 的应用方式。后续可考虑将 scale 计算统一到模型配置层，避免重复代码。

功能与动机

PR body 指出："The prefill backend refactor in f3fef12350 (#32623) started constructing MLA prefill backends with model_config.get_head_size() **-0.5. For DeepSeek-Coder-V2-Lite, that head size is the 576-dim latent KV cache, not the 192-dim query/key attention head. It also missed the DeepSeek YaRN mscale correction applied by the model attention module." 该问题导致 ROCm CI 中 DeepSeek GSM8K 测试失败。

实现拆解

1. 新增 get_mla_prefill_scale 函数（vllm/model_executor/layers/attention/mla_attention.py:1333-1359）：基于 mla_dims 计算 qk_head_dim，然后判断模型类型（通过 compress_ratios 属性区分 DeepSeek V4），接着从 rope_parameters 中提取 YaRN 相关参数并应用 mscale 校正。
2. 修改 MLAAttentionImpl.__init__（同一文件 :1562）：原硬编码 scale=self.model_config.get_head_size() **-0.5 替换为 scale=get_mla_prefill_scale(self.model_config)，并将 yarn_get_mscale 导入。
3. 新增单元测试（tests/v1/attention/test_mla_prefill_selector.py:61-114）：TestMLAPrefillScale 类包含三个测试用例，分别验证 DeepSeek V2 风格、带 YaRN mscale 的 V2/V3、以及 V4 风格（不应用 mscale）的 scale 计算。
4. 调整后端正确性测试（tests/v1/attention/test_mla_backends.py：789-790, 907, 943）：将原有的单一 scale 变量拆分为 decode_scale（保持原缩放）和 prefill_scale（调用 get_mla_prefill_scale），并分别在 decode 和 prefill 的 SDPA 调用中使用对应 scale。

关键文件：

- vllm/model_executor/layers/attention/mla_attention.py（模块 注意力层；类别 source；类型 core-logic；符号 get_mla_prefill_scale）：核心源码文件，新增

get_mla_prefill_scale 函数并修改 MLAAttentionImpl.__init__ 中的 scale 计算。

- tests/v1/attention/test_mla_prefill_selector.py（模块 预填选择器；类别 test；类型 test-coverage；符号 TestMLAPrefillScale, test_uses_qk_head_dim_for_deepseek_v2_style_mla, test_applies_deepseek_yarn_mscale, test_deepseek_v4_style_mla_does_not_apply_yarn_mscale）：新增 TestMLAPrefillScale 测试类，覆盖三种典型配置的 scale 计算，确保修复正确性。
- tests/v1/attention/test_mla_backends.py（模块 MLA 后端；类别 test；类型 test-coverage）：调整后端正确性测试，区分 decode 和 prefill 的 scale，引入 get_mla_prefill_scale。

关键符号：get_mla_prefill_scale

关键源码片段

vllm/model_executor/layers/attention/mla_attention.py

核心源码文件，新增 get_mla_prefill_scale 函数并修改 MLAAttentionImpl.__init__ 中的 scale 计算。

```
def get_mla_prefill_scale(model_config: ModelConfig) -> float:
    hf_text_config = model_config.hf_text_config
    mla_dims = get_mla_dims(model_config)
    # query/key attention head 维度由 nope 部分和 rope 部分组成
    qk_head_dim = mla_dims.qk_nope_head_dim + mla_dims.qk_rope_head_dim
    scale = qk_head_dim ** -0.5

    # DeepSeek V4 的 attention 路径禁用了 YaRN mscale,
    # 而 DeepSeek V2/V3 在构建 MLA attention 模块时会应用相同的 mscale 校正
    if hasattr(hf_text_config, "compress_ratios"):
        return scale

    rope_parameters = getattr(hf_text_config, "rope_parameters", None)
    if rope_parameters is None:
        rope_parameters = getattr(hf_text_config, "rope_scaling", None)

    # 如果没有 rope 参数，直接返回基础 scale
    if rope_parameters is None:
        return scale

    # 检查 rope_type 和 apply_yarn_scaling 标志
    rope_type = rope_parameters.get("rope_type", rope_parameters.get("type"))
    apply_yarn_scaling = rope_parameters.get("apply_yarn_scaling", True)
    # 仅当 rope_type 不是 "default" 且 yarn scaling 启用时应用 mscale
    if rope_type != "default" and apply_yarn_scaling:
        mscale_all_dim = rope_parameters.get("mscale_all_dim", False)
        scaling_factor = rope_parameters["factor"]
        mscale = yarn_get_mscale(float(scaling_factor), float(mscale_all_dim))
        scale *= mscale * mscale
```

```
    return scale
```

tests/v1/attention/test_mla_prefill_selector.py

新增 TestMLAPrefillScale 测试类，覆盖三种典型配置的 scale 计算，确保修复正确性。

```
class TestMLAPrefillScale:
    """Tests for the MLA prefill softmax scale."""

    def test_uses_qk_head_dim_for_deepseek_v2_style_mla(self):
        # DeepSeek V2 风格配置：使用 qk_nope_head_dim + qk_rope_head_dim
        model_config = SimpleNamespace(
            hf_text_config=SimpleNamespace(
                q_lora_rank=None,
                kv_lora_rank=512,
                qk_nope_head_dim=128,
                qk_rope_head_dim=64,
                v_head_dim=128,
                rope_parameters={"rope_type": "default"},
            )
        )
        # scale 应为 192**-0.5（128+64=192）
        assert get_mla_prefill_scale(model_config) == pytest.approx(192**-0.5)

    def test_applies_deepseek_yarn_mscale(self):
        # DeepSeek V2/V3 使用 YaRN 时，scale 需乘 mscale^2
        model_config = SimpleNamespace(
            hf_text_config=SimpleNamespace(
                q_lora_rank=None,
                kv_lora_rank=512,
                qk_nope_head_dim=128,
                qk_rope_head_dim=64,
                v_head_dim=128,
                rope_parameters={
                    "rope_type": "yarn",
                    "factor": 40,
                    "mscale_all_dim": 0.707,
                },
            )
        )
        mscale = yarn_get_mscale(40, 0.707)
        assert get_mla_prefill_scale(model_config) == pytest.approx(
            192**-0.5 * mscale * mscale
        )

    def test_deepseek_v4_style_mla_does_not_apply_yarn_mscale(self):
        # DeepSeek V4 配置（含 compress_ratios）应忽略 YaRN mscale
        model_config = SimpleNamespace(
            hf_text_config=SimpleNamespace(
                compress_ratios=[4],
```

```
            q_lora_rank=1536,
            head_dim=128,
            qk_rope_head_dim=64,
            rope_parameters={
                "rope_type": "yarn",
                "factor": 40,
                "mscale_all_dim": 0.707,
            },
        )
    )
    # scale 应为 128**-0.5（只使用 head_dim，而非 qk 维度）
    assert get_mla_prefill_scale(model_config) == pytest.approx(128**-0.5)
```

评论区精华

- 安全访问 factor 键：gemini-code-assist 指出直接使用 rope_parameters["factor"] 可能引发 KeyError，建议使用 .get() 或备用键。作者 AndreasKaratzas 回复 "I think this is too defensive"，认为该配置在 DeepSeek 模型中必然存在，未采纳。
- 代码重复担忧：MatthewBonanni 在 approval 中评论 "I don't love the duplicated code in get_mla_prefill_scale"，但认为当前修复可先让 CI 通过，后续应尽快提供更好的方案。mgoin 和 MatthewBonanni 均 approve。
  - rope_parameters["factor"] 安全访问 (correctness): 未修改，维持直接索引。
  - 重复代码触发更好的架构重构 (design): 先合并，后续改进。

风险与影响

- 风险：回归风险：低。修复仅影响 MLA prefill 的 scale 计算，且函数与现有模型配置解耦。测试覆盖了 DeepSeek V2/V3/V4 三种风格。性能风险：无，scale 计算为纯 Python 预处理，不涉及运行时 kernel。兼容性风险：若其他模型意外依赖了旧 scale（如非 DeepSeek 但使用 MLA 的模型），修复会改变其数值行为，但此类模型在当前代码库中不存在。配置安全风险：直接索引 rope_parameters["factor"] 在异常配置下可能 KeyError，但实际 DeepSeek 配置必然包含该键。
- 影响：用户影响：使用 DeepSeek 系列模型（V2/V3/V4）且启用 MLA prefill 的用户将得到正确的 prefill 输出，修复了 GSM8K 评测准确率下降的问题。系统影响：新增的 get_mla_prefill_scale 函数成为 MLA prefill scale 的唯一来源，便于后续维护。测试影响：新增 3 个单元测试和 1 个集成测试调整，提高了置信度。
- 风险标记：核心路径变更，配置安全访问，跨模型变体条件

关联脉络

- PR #32623 prefill backend refactor: 引入 scale 计算错误，本 PR 修复该回归。