PR #25390 完整报告

sgl-project/sglang

[AMD] Enable shared-experts fusion with new KIMI-K2.5-MXFP4 model.

合并时间：2026-05-18 16:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25390

执行摘要

- 一句话：AMD 启用 KIMI-K2.5-MXFP4 共享专家融合
- 推荐动作：该 PR 值得精读，尤其是学习如何通过最小改动适配新模型架构。重点关注：
  - quark.py 中名称映射的扩展模式，可推广到其他多模态模型。
  - 允许列表模式，为未来支持更多专家数量提供了范例。
  - 但需注意已识别的引用安全问题，建议后续修复。

功能与动机

使 Kimi-K2.5-MXFP4 (Quark MXFP4 检查点) 能在 AMD MI3xx 上实际加载，并受益于 DeepSeek-V3/R1 已使用的共享专家融合优化。

实现拆解

1. quark.py – 扩展排除层名称映射 - 修改 apply_weight_name_mapper 方法：将 HuggingFace 名称映射后，对每个以 language_model. 开头的名称，额外保留去掉该前缀的版本。 - 使用 list(dict.fromkeys(expanded)) 去重并保持顺序，避免多模态包装检查点中的命名差异导致排除层无法匹配。 - 对非多模态检查点无影响（前缀不匹配则不生成额外条目）。
2. deepseek_v2.py – 传播 fused-module 映射 - 在 DeepseekV2ForCausalLM.__init__ 中，将模型自身的 packed_modules_mapping 赋值给 quant_config.packed_modules_mapping（如果存在该属性）。 - 使量化配置能访问 fused-module 映射，从而在 should_ignore_layer 等排除检查中正确解引用融合后的名称。
3. deepseek_v2.py – 扩展允许的专家数量 - 在 determine_num_fused_shared_experts 中，将 n_routed_experts 的硬编码检查从 != 256 改为 not in (256, 384)。 - 新增注释说明：256 对应 DeepSeek-V3/R1，384 对应 Kimi-K2.5。 - 其他未经验证的值仍走安全禁用路径，确保不影响未知配置。

关键文件：

- python/sglang/srt/models/deepseek_v2.py（模块 模型定义；类别 source；类型 data-contract；符号 DeepseekV2ForCausalLM.init, determine_num_fused_shared_experts）：核心模型文件，修改了 shared-experts fusion 的准入条件并传播 packed_modules_mapping 给量化配置

- python/sglang/srt/layers/quantization/quark/quark.py（模块 量化层；类别 source；类型 core-logic；符号 apply_weight_name_mapper）：量化配置层，扩展排除层名称映射以支持多模态包装的命名约定

关键符号：DeepseekV2ForCausalLM.init, determine_num_fused_shared_experts, QuarkConfig.apply_weight_name_mapper

关键源码片段

python/sglang/srt/models/deepseek_v2.py

核心模型文件，修改了 shared-experts fusion 的准入条件并传播 packed_modules_mapping 给量化配置

```
# 在 __init__ 中，将 fused-module 映射传播给量化配置
# 这样量化配置在排除检查时可以正确解引用融合后的层名
if quant_config is not None and hasattr(quant_config, "packed_modules_mapping"):
    quant_config.packed_modules_mapping = self.packed_modules_mapping

# 在 determine_num_fused_shared_experts 中，扩展允许的 n_routed_experts
elif (
    self.config.architectures[0] != architecture
    # Allow-list of n_routed_experts values that have been validated
    # for shared-experts fusion under this code path. Currently:
    # 256 -> DeepSeek-V3 / R1
    # 384 -> Kimi-K2.5 (text_config wraps DeepseekV3ForCausalLM)
    or self.config.n_routed_experts not in (256, 384)
    or self.config.n_shared_experts != 1
):
    disable_reason = "Config does not support fused shared expert(s)."
```

python/sglang/srt/layers/quantization/quark/quark.py

量化配置层，扩展排除层名称映射以支持多模态包装的命名约定

```
# 在 apply_weight_name_mapper 中，映射后扩展排除层列表
# 多模态检查点的层名可能以 'language_model.' 为前缀，也可能没有
# 因此保留两种形式，确保排除检查总能命中
def apply_weight_name_mapper(self, hf_to_sglang_mapper):
    mapped = hf_to_sglang_mapper.apply_list(self.exclude_layers)
    expanded = []
    for name in mapped:
        expanded.append(name)
        if name.startswith("language_model."):
            expanded.append(name.removeprefix("language_model."))
    self.exclude_layers = list(dict.fromkeys(expanded)) # 去重并保持顺序
```

评论区精华

主要讨论点：引用安全问题

- gemini-code-assist[bot] 指出：直接赋值 quant_config.packed_modules_mapping = self.packed_modules_mapping 创建了对类级字典的引用，后续对 quant_config.packed_modules_mapping 的修改会意外修改 DeepseekV2ForCausalLM.packed_modules_mapping 类属性，可能影响其他模型实例。建议使用 .update() 合并。
- 结论：未在 PR 中解决该问题，但 PR 已获 HaiShaw 批准并合并。潜在风险仍在。
  - 直接赋值 packed_modules_mapping 的引用安全问题 (correctness): 未采纳建议，PR 已合并，风险未修复。

风险与影响

- 风险：
  - 引用共享风险：deepseek_v2.py 中直接赋值 quant_config.packed_modules_mapping = self.packed_modules_mapping，若量化配置后续修改该字典，将意外影响类级属性，可能导致多实例间的状态污染。
  - 回归风险低：对 DeepSeek-V3/R1 而言，n_routed_experts 为 256 仍在允许集合中，行为不变。
  - 缺少测试覆盖：未添加针对 Kimi-K2.5 或新 n_routed_experts 值的单元测试，回归风险依赖人工验证。
- 影响：
  - 用户：AMD MI3xx 用户现可加载 Kimi-K2.5-MXFP4 并享受共享专家融合优化，吞吐量提升约 15%，延迟降低约 13%。
  - 系统：仅影响使用了 Quark 量化且模型为 DeepseekV2ForCausalLM 架构的多模态检查点。
  - 团队：为后续支持其他专家数量（如 512）提供了清晰的扩展模式，只需在元组中添加新值。
  - 风险标记：引用共享风险，缺少测试覆盖

关联脉络

- PR #22822 [Refactor] Refactor DeepEP dispatcher: 同为模型适配相关，涉及 DeepSeek 系列模型的量化配置优化。