PR #25257 完整报告

sgl-project/sglang

[NPU] Support model DeepSeek-OCR and DeepSeek-OCR-2

合并时间：2026-05-21 15:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25257

执行摘要

- 一句话：NPU 支持 DeepSeek-OCR 系列模型
- 推荐动作：值得快速合入，解决了 NPU 平台特定模型的功能阻塞。设计上采用条件导入而非运行时派发，是合理的权衡。后续建议在 CI 中增加 NPU DeepSeek-OCR 的回归测试。

功能与动机

PR body 说明需要支持 DeepSeek-OCR 和 DeepSeek-OCR-2 模型在 NPU 平台上运行。Issue 评论中 zhsurpass 指出图片请求时 GPU 图像解码返回张量导致失败，关联 issue #24699 和 fix PR #24701。

实现拆解

1. 导入依赖调整：在 python/sglang/srt/models/deepseek.py 中，新增 is_npu 工具函数导入，并定义模块级布尔变量 _is_npu。
2. 条件导入 fused_moe：根据 _is_npu 的值，在模块加载时选择导入 fused_moe_npu（来自 sglang.srt.hardware_backend.npu.quantization.fused_moe_method_npu）或默认的 Triton fused_moe。原模块级导入 from sglang.srt.layers.moe.moe_runner.triton_utils import fused_moe 被移除。
3. forward 调用简化：在 DeepseekMoE.forward() 中，原来调用 fused_moe.fused_moe(...) 的写法简化为直接调用 fused_moe(...)，因为现在导入的是函数对象而非模块。

关键文件：

- python/sglang/srt/models/deepseek.py（模块 模型层；类别 source；类型 core-logic）：核心变更文件，修改了 DeepSeekMoE 在 NPU 上的 fused_moe 导入逻辑和调用方式。

关键符号：DeepseekMoE.forward

关键源码片段

python/sglang/srt/models/deepseek.py

核心变更文件，修改了 DeepSeekMoE 在 NPU 上的 fused_moe 导入逻辑和调用方式。

```
# python/sglang/srt/models/deepseek.py
```

```
# 在文件顶部导入 is_npu 工具函数
from sglang.srt.utils import add_prefix, cpu_has_amx_support, is_cpu, is_npu

_is_cpu = is_cpu()
_is_npu = is_npu() # 新增：判断当前是否为 NPU 平台

# 根据平台条件导入 fused_moe 实现
# NPU 使用专用的 fused_moe_npu，其他平台使用默认的 Triton fused_moe
if _is_npu:
    from sglang.srt.hardware_backend.npu.quantization.fused_moe_method_npu import (
        fused_moe_npu as fused_moe,
    )
else:
    from sglang.srt.layers.moe.moe_runner.triton_utils.fused_moe import fused_moe

class DeepseekMoE(nn.Module):
    # ... 省略初始化代码 ...

    def forward(self, hidden_states: torch.Tensor) -> torch.Tensor:
        # ... 省略前处理 ...
        if _is_cpu and _is_cpu_amx_available:
            # CPU AMX 路径保持不变
            final_hidden_states = torch.ops.sgl_kernel.fused_experts_cpu(...)
        else:
            # 简化调用：直接使用 fused_moe（根据平台已正确导入）
            final_hidden_states = fused_moe(
                hidden_states,
                w1=self.w1,
                w2=self.w2,
                topk_output=topk_output,
                moe_runner_config=MoeRunnerConfig(inplace=True),
            )
        # ... 省略后处理 ...
```

评论区精华

讨论较少，主要涉及 CI 重试。gemini-code-assist[bot] 的 review 确认变更合理，无反馈。sglang-npu-bot 批准合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。条件导入在模块加载时执行，不影响运行时性能。非 NPU 路径行为不变。潜在风险：NPU 上的 fused_moe_npu 实现需与模型权重格式兼容，若存在差异可能导致精度问题；但 PR 附件中的精度测试结果（OmniDocBench）显示通过。
- 影响：影响范围小，仅影响 NPU 平台上 DeepSeek-OCR 和 DeepSeek-OCR-2 模型的推理。对其他平台或模型无影响。

- 风险标记：缺少单元测试，依赖特定硬件的 fused_moe 实现

关联脉络

- PR #24701 Fix image decoding for DeepSeek-OCR on NPU: 关联 issue #24699 提到的修复 PR，与本 PR 共同解决 DeepSeek-OCR 在 NPU 上的问题。