PR #43891 完整报告

vllm-project/vllm

[Model Refactoring] Remove unncessary torch op registration for DSv4

合并时间：2026-05-29 05:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43891

执行摘要

- 一句话：移除 DSv4 中不必要的 torch op 注册
- 推荐动作：对于关注 DSv4 模型开发的同学，值得阅读此 PR 以了解如何清理 torch.compile 依赖。对于其他模型开发者，可作为简单的重构样例。

功能与动机

PR body 明确指出："Simple cleanup: Now that DSv4 does not rely on torch compile, we don't need to wrap the deepgemm kernel calls with the torch op registration." 因此解除不再需要的 op 注册，简化代码。

实现拆解

1. 在 attention.py 中：移除 from vllm.utils.torch_utils import direct_register_custom_op 导入；删除 deepseek_v4_attention_fake、deepseek_v4_fp8_einsum、deepseek_v4_fp8_einsum_fake 三个函数及其 direct_register_custom_op 注册；将 torch.ops.vllm.deepseek_v4_attention 调用改为直接调用 deepseek_v4_attention；将 torch.ops.vllm.deepseek_v4_fp8_einsum 调用改为直接调用 fp8_einsum，并调整参数传递方式（使用元组传递张量与 scale，添加 recipe 参数）。
2. 在 nvidia/model.py 中：移除 from vllm.utils.torch_utils import direct_register_custom_op 和 from vllm.forward_context import get_forward_context 导入；删除 _deepseek_v4_mega_moe_experts_op 和 _deepseek_v4_mega_moe_experts_op_fake 函数及其注册；将 torch.ops.vllm.deepseek_v4_mega_moe_experts 调用替换为直接调用 self._run_mega_moe 方法；同时移除 __init__ 中不再需要的 static_forward_context 注册代码。
3. 总计净减少约 110 行代码，不改变运行行为，无需新增测试。

关键文件：

- vllm/models/deepseek_v4/attention.py（模块 注意力；类别 source；类型 core-logic；符号 deepseek_v4_attention_fake, deepseek_v4_fp8_einsum, deepseek_v4_fp8_einsum_fake, deepseek_v4_attention）：核心注意力层代码，移除了 deepseek_v4_attention 和 deepseek_v4_fp8_einsum 的 torch op 注册，改为直接调用函数。

- vllm/models/deepseek_v4/nvidia/model.py（模块 MoE；类别 source；类型 core-logic；符号 _deepseek_v4_mega_moe_experts_op, _deepseek_v4_mega_moe_experts_op_fake）：MoE 层代码，移除了 _deepseek_v4_mega_moe_experts_op 的 torch op 注册，改为直接调用 self._run_mega_moe。

关键符号：deepseek_v4_attention_fake, deepseek_v4_fp8_einsum, deepseek_v4_fp8_einsum_fake, _deepseek_v4_mega_moe_experts_op, _deepseek_v4_mega_moe_experts_op_fake

关键源码片段

vllm/models/deepseek_v4/attention.py

核心注意力层代码，移除了 deepseek_v4_attention 和 deepseek_v4_fp8_einsum 的 torch op 注册，改为直接调用函数。

```
# 前向函数中直接调用 `deepseek_v4_attention`，不再通过 torch.op
# @eager_break_during_capture: 该调用在 CUDA graph 捕获的图段之间以 eager 模式运行
deepseek_v4_attention(
    hidden_states,
    positions,
    o_padded,
    self.layer_name,
)
o = o_padded[:, : self.n_local_heads, :]

# O projection: 使用 `fp8_einsum` 直接计算，不再通过 torch.op
# 注意：recipe 参数现在直接传递，不需要 tuple(recipe) 转换
z = torch.empty(
    (num_tokens, self.n_local_groups, self.o_lora_rank),
    device=o.device,
    dtype=torch.bfloat16,
)
fp8_einsum(
    "bhr,hdr->bhd",
    (o_fp8, o_scale),
    (wo_a_fp8, wo_a_scale),
    z,
    recipe=self._einsum_recipe,
)
return self.wo_b(z.flatten(1))
```

vllm/models/deepseek_v4/nvidia/model.py

MoE 层代码，移除了 _deepseek_v4_mega_moe_experts_op 的 torch op 注册，改为直接调用 self._run_mega_moe。

```
# 在 DeepseekV4MegaMoEExperts.forward 中：
# 移除 torch.ops.vllm.deepseek_v4_mega_moe_experts 调用，直接调用 self._run_mega_moe
y = torch.empty_like(hidden_states, dtype=torch.bfloat16)
self._run_mega_moe(
```

```
    hidden_states,
    topk_weights,
    topk_ids,
    y,
    activation_clamp,
    fast_math,
)
```

评论区精华

该 PR 无 review 评论，仅获得 zyongye 的 approve。无实质性讨论。

- 暂无高价值评论线程

风险与影响

- 风险：变更非常清晰：移除已经不再需要的包装层，直接调用底层函数。风险极低，因为功能完全等价，且不涉及逻辑变更。但未增加测试覆盖，若有未来回归，需要依赖现有模型测试。
- 影响：仅影响 DSv4 模型的注意力层和 MoE 层，对其他模型无影响。代码可维护性提升，开发者不再需要维护 fake 实现和 op 注册。
- 风险标记：低风险，无新增测试

关联脉络

- 暂无明显关联 PR