PR #25569 完整报告

sgl-project/sglang

Add DeepSeekV4 fused MoE Triton autotune support

合并时间：2026-05-18 21:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25569

执行摘要

- 一句话：为 DeepSeekV4 添加 fused MoE Triton autotune 支持
- 推荐动作：PR 改动简洁清晰，适合作为支持新模型架构调优的参考模板。建议关注 swiglu_limit 值的通用性，未来可考虑从模型配置中自动推导。

功能与动机

提升 DeepSeekV4 模型在 fused MoE Triton 核上的推理性能，通过 autotune 自动搜索最优配置。PR body 明确指出 "Add DeepSeekV4 fused MoE Triton autotune support to improve inference performance"。

实现拆解

1. 导入新增：在 tuning_fused_moe_triton.py 中添加 get_config 导入，用于运行时从 HuggingFace 加载模型配置。
2. 架构检测：在 run() 函数内部通过 get_config 获取 architectures[0] 并判断是否为 DeepseekV4ForCausalLM。
3. 参数调整：若检测到 DeepSeekV4，则设置 MoeRunnerConfig.swiglu_limit=10.0；否则保持默认 None。
4. 架构注册：在 common_utils.py 的 DeepSeek 系列架构列表中加入 DeepseekV4ForCausalLM，确保 get_model_config 等函数能正确解析其 MoE 参数。
5. 无测试配套：本次变更仅涉及 benchmark 工具目录，未引入新的单元测试。

关键文件：

- benchmark/kernels/fused_moe_triton/tuning_fused_moe_triton.py（模块 调优脚本；类别 source；类型 dependency-wiring）：核心改动：添加导入和架构检测逻辑，根据 DeepSeekV4 设置 swiglu_limit 参数以适配其 fused MoE 实现。
- benchmark/kernels/fused_moe_triton/common_utils.py（模块 配置工具；类别 source；类型 core-logic）：配置扩展：在 DeepSeek 架构支持列表中加入 DeepseekV4ForCausalLM，使其能通过架构检查并获取正确的模型参数用于调优。

关键符号：未识别

关键源码片段

benchmark/kernels/fused_moe_triton/tuning_fused_moe_triton.py

核心改动：添加导入和架构检测逻辑，根据 DeepSeekV4 设置 swiglu_limit 参数以适配其 fused MoE 实现。

```
def run():
    # 从 HuggingFace 加载模型配置，获取架构名称
    model_config = get_config(args.model, trust_remote_code=True)
    architecture = model_config.architectures[0]
    # 判断是否为 DeepSeekV4，用于设置特定的 swiglu_limit
    is_dsv4 = architecture == "DeepseekV4ForCausalLM"
    moe_runner_config = MoeRunnerConfig(
        inplace=True,
        # DeepSeekV4 需要较低的 swiglu_limit 以获得最佳性能
        swiglu_limit=10.0 if is_dsv4 else None,
    )

    with override_config(config):
        fused_moe(
            x,
            w1,
            w2,
            topk_output,
            moe_runner_config=moe_runner_config,
            use_fp8_w8a8=use_fp8_w8a8,
            use_int8_w8a8=use_int8_w8a8,
            use_int8_w8a16=use_int8_w8a16,
            use_int4_w4a16=use_int4_w4a16,
            w1_scale=w1_scale,
            w2_scale=w2_scale,
            a1_scale=a1_scale,
            a2_scale=a2_scale,
            per_channel_quant=per_channel_quant,
            block_shape=block_shape,
        )
```

benchmark/kernels/fused_moe_triton/common_utils.py

配置扩展：在 DeepSeek 架构支持列表中加入 DeepseekV4ForCausalLM，使其能通过架构检查并获取正确的模型参数用于调优。

```
    # 在 get_model_config 函数中，将 DeepseekV4ForCausalLM 添加到 DeepSeek 系列架构列表
    elif architecture in [
        "DeepseekV2ForCausalLM",
        "DeepseekV3ForCausalLM",
        "DeepseekV32ForCausalLM",
        "DeepseekV4ForCausalLM", # 新增：支持 DeepSeekV4
        "Glm4MoeForCausalLM",
        "GlmMoeDsaForCausalLM",
        "MistralLarge3ForCausalLM",
    ]:
```

```
E = (config.n_routed_experts // ep_size) + (
    0
    if disable_shared_experts_fusion
    or architecture
    not in [
        "DeepseekV3ForCausalLM",
        "DeepseekV32ForCausalLM",
        "Glm4MoeForCausalLM",
        "GlmMoeDsaForCausalLM",
        "MistralLarge3ForCausalLM",
    ]
    else 1
)
topk = config.num_experts_per_tok + (
    0 if disable_shared_experts_fusion or topk_ids_dir is None else 1
)
intermediate_size = config.moe_intermediate_size
```

评论区精华

无 review 评论，仅 BBuf 直接批准。未发现技术争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：改动仅限 benchmark 工具，不涉及生产推理路径。但需注意：
  - swiglu_limit=10.0 是经验值，不同配置可能需要重新调优。
  - get_config 加载模型配置可能依赖 HuggingFace 网络，建议在离线环境缓存。
  - 其他 DeepSeek 架构（如 V2/V3）的行为不受影响，因为 is_dsv4 仅对 V4 生效。
  - 影响：影响范围有限，仅提升 DeepSeekV4 在 benchmark 调优中的效率。对系统性能无直接负面影响，但为后续将调优结果集成到生产环境提供了基础。
  - 风险标记：benchmark-only, small-change

关联脉络

- 暂无明显关联 PR