PR #1977 完整报告

THUDM/slime

[docker] fix GLM4.7 flash for sglang v0.5.12

合并时间：2026-05-29 17:21

原文链接：http://prhub.com.cn/THUDM/slime/pull/1977

执行摘要

- 一句话：修复 GLM4.7 在 sglang v0.5.12 下的运行问题
- 推荐动作：建议项目组其他成员了解变更，尤其是需要维护 GLM4 系列模型的同学。可通过此 PR 学习 patch 中的条件模型初始化技巧。

功能与动机

GLM4.7 模型（含 MOE 和视觉变体）在升级到 sglang v0.5.12 后出现兼容性问题，主要是 encoder_only 模式和 shared expert 的 TP1 判断逻辑需要适配。作者通过 patch 形式修复这些模型定义，确保 GLM4 系列在 slime 框架下正常运行。

实现拆解

1. glm4v_moe.py model 条件创建：在 Glm4vMoeForConditionalGeneration.__init__ 中，将原先无条件创建 self.model 改为 if not self.config.encoder_only: 才创建，避免 encoder 模式下创建错误的 language model。
2. glm4v_moe.py lm_head 条件创建：在创建 model 后，增加 if self.pp_group.is_last_rank: 判断，根据 tie_word_embeddings 和 world_size 决定使用 embed_tokens 还是 ParallelLMHead，确保正确初始化 lm_head。
3. glm4_moe_lite.py shared expert TP1 逻辑重构：提取 shared_expert_use_tp1 逻辑为独立变量，统一 get_moe_a2a_backend().is_deepep()、is_mooncake() 和 should_use_flashinfer_cutlass_moe_fp4_allgather() 判断，并增加 _shared_expert_tp1 属性记录结果，后续可能用于决策参考。
4. glm4_moe_lite.py 废弃注释与初始化：取消 shared_experts_weight_block_size 的注释，显式初始化为 None；增加 _shared_expert_tp1 属性初始化为 False。

关键文件：

- docker/patch/latest/sglang.patch（模块 补丁脚本；类别 infra；类型 core-logic）：所有变更均在此 patch 文件中，覆盖 glm4v_moe.py 和 glm4_moe_lite.py 的模型初始化与 shared expert 逻辑。

关键符号：未识别

关键源码片段

docker/patch/latest/sglang.patch

所有变更均在此 patch 文件中，覆盖 glm4v_moe.py 和 glm4_moe_lite.py 的模型初始化与 shared expert 逻辑。

```
# docker/patch/latest/sglang.patch 中 glm4v_moe.py 变更节选
classGlm4vMoeForConditionalGeneration(Glm4vForConditionalGeneration):
def__init__(self, config, quant_config, prefix):        # ... 原有逻辑 ...        # 仅在非 encoder_only 模式下创建 language model        if not self.config.encoder_only:
self.model = Glm4MoeModel(                config, quant_config,
prefix=add_prefix("language_model", prefix),            )            # 仅在最后 rank 上创建 lm_head            if self.pp_group.is_last_rank:                if self.pp_group.world_size == 1 and self.config.tie_word_embeddings:                    self.lm_head = self.model.embed_tokens # 权重绑定                else:                    self.lm_head = ParallelLMHead(                        config.vocab_size, config.hidden_size,
    quant_config=quant_config,                        prefix=add_prefix("lm_head", prefix),
                use_attn_tp_group=get_global_server_args().enable_dp_lm_head,
        ) # docker/patch/latest/sglang.patch 中 glm4_moe_lite.py 变更节选
classGlm4MoeLiteSparseMoeBlock(DeepseekV2MoE):    def__init__(self, config, quant_config, prefix):        # ... 原有初始化 ...
self.shared_experts_weight_block_size = None # 取消注释，显式初始化
self._shared_expert_tp1 = False        if config.n_shared_experts is not None and self.num_fused_shared_experts == 0:            intermediate_size = config.moe_intermediate_size * config.n_shared_experts            # 提前判定 shared expert 是否需要 TP1            shared_expert_use_tp1 = (
get_moe_a2a_backend().is_deepep()                or
get_moe_a2a_backend().is_mooncake()                or
should_use_flashinfer_cutlass_moe_fp4_allgather()            )
self.shared_experts = Glm4MoeLiteMLP(                hidden_size=config.hidden_size,
        intermediate_size=intermediate_size,                # ... 其他参数 ...
**(dict(tp_rank=0, tp_size=1) if shared_expert_use_tp1 else {}),            )
self._shared_expert_tp1 = shared_expert_use_tp1 # 记录判定结果
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。补丁仅影响 GLM4 系列模型（glm4v_moe 和 glm4_moe_lite），且改动集中在模型初始化和 shared expert 配置。可能的风险：encoder_only 分支未覆盖所有初始化路径，或 shared_expert_use_tp1 变量提取引入回归（但逻辑等价，概率低）。未关联其他模块。

- 影响：直接影响：GLM4.7 模型（含 GLM4v-MOE 和 GLM4-MoE-Lite）在 sglang v0.5.12 上可用。间接受益：未来类似模型（如 DeepSeek-V2 MoE 变体）可能复用 shared_expert_tp1 逻辑。无用户可见的 API 变更。
- 风险标记：仅补丁级变更，缺少测试覆盖

关联脉络

- PR #1978 [docker] Fix qwen3 30B + deepep with sglang 0.5.12: 同为 sglang v0.5.12 的 docker 补丁，聚焦不同模型问题，展示持续兼容性维护。
- PR #1973 [docker] fix patch: 最近的 sglang patch 修复，与当前 PR 同属 docker 补丁系列。
- PR #1972 [docker] fix mooncake offload in sglang v0.5.12: 同为 docker 补丁，修复 mooncake offload，与当前 PR 的 mooncake 条件关联。