PR #38278 完整报告

vllm-project/vllm

[Model] Use AutoWeightsLoader for InternLM2

合并时间：2026-05-26 18:39

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38278

执行摘要

- 一句话：InternLM2 迁移至 AutoWeightsLoader
- 推荐动作：该 PR 设计清晰，遵循已有模式，适合作为参考样本帮助其他模型的迁移工作。建议阅读 InternLM2Model.load_weights 和 InternLM2ForCausalLM.load_weights 的变更以理解 AutoWeightsLoader 的使用方式。

功能与动机

作为 Issue #15697（为所有模型使用 AutoWeightsLoader 实现复合模型加载）的一部分，将 InternLM2 的权重加载标准化，使 InternLM2ForRewardModel 能自动继承 InternLM2Model 的 load_weights，并利用 AutoWeightsLoader 跳过 StageMissingLayer 模块、直接路由 v_head 权重。

实现拆解

1. 在 InternLM2Model 类中新增 load_weights 方法（vllm/model_executor/models/internlm2.py）：将与原始 InternLM2ForCausalLM.load_weights 相同的逻辑（stacked_params_mapping、参数遍历、shard_id 分发）移至 InternLM2Model，保持对 merged linear layers 的兼容性。
2. 将 InternLM2ForCausalLM.load_weights 重写为使用 AutoWeightsLoader 的简单委托：创建 AutoWeightsLoader 实例，传入 self 和 skip_prefixes（当 tie_word_embeddings 启用时跳过 'output.'），并调用 loader.load_weights，返回加载的参数集合。
3. 更新 import：在文件头部添加 AutoWeightsLoader 导入（from .utils import AutoWeightsLoader）。
4. 移除重复逻辑：删除原先同样位于 InternLM2ForCausalLM 中的手动权重加载代码，避免重复。

关键文件：

- vllm/model_executor/models/internlm2.py（模块 模型执行器；类别 source；类型 data-contract；符号 load_weights）：核心变更文件：将 load_weights 逻辑从 InternLM2ForCausalLM 迁移到 InternLM2Model，并引入 AutoWeightsLoader 委托。同时新增 import 和调整权重加载路径，是 PR 的唯一修改文件。

关键符号：load_weights

关键源码片段

vllm/model_executor/models/internlm2.py

核心变更文件：将 load_weights 逻辑从 InternLM2ForCausalLM 迁移到 InternLM2Model，并引入 AutoWeightsLoader 委托。同时新增 import 和调整权重加载路径，是 PR 的唯一修改文件。

```
# 位于 InternLM2Model 类内（新增的 load_weights）
def load_weights(self, weights: Iterable[tuple[str, torch.Tensor]]) -> set[str]:
    stacked_params_mapping = [
        # (param_name, shard_name, shard_id)
        ("gate_up_proj", "w1", 0),
        ("gate_up_proj", "w3", 1),
    ]
    params_dict = dict(self.named_parameters())
    loaded_params: set[str] = set()
    for name, loaded_weight in weights:
        if "rotary_emb.inv_freq" in name:
            continue
        for param_name, weight_name, shard_id in stacked_params_mapping:
            if weight_name not in name:
                continue
            name = name.replace(weight_name, param_name)
            # 跳过 GPTQ 模型的额外 bias
            if name.endswith(".bias") and name not in params_dict:
                continue
            if is_pp_missing_parameter(name, self):
                continue
            param = params_dict[name]
            weight_loader = param.weight_loader
            weight_loader(param, loaded_weight, shard_id)
            break
        else:
            # 非 stacked 参数的处理
            if name.endswith(".bias") and name not in params_dict:
                continue
            if is_pp_missing_parameter(name, self):
                continue
            param = params_dict[name]
            weight_loader = getattr(param, "weight_loader", default_weight_loader)
            weight_loader(param, loaded_weight)
            loaded_params.add(name)
    return loaded_params

# 位于 InternLM2ForCausalLM 类内（重构后的 load_weights）
def load_weights(self, weights: Iterable[tuple[str, torch.Tensor]]) -> set[str]:
    loader = AutoWeightsLoader(
        self,
```

```
            skip_prefixes=(["output."] if self.config.tie_word_embeddings else None),
        )
        return loader.load_weights(weights)
```

评论区精华

gemini-code-assist[bot] 指出 InternLM2Model.load_weights 仍为手动实现，与“使用 AutoWeightsLoader”的目标矛盾。作者 javierdejesusda 回应这是标准模式：AutoWeightsLoader 不处理 shard_id 分发，因此内部 *Model 类需保留手动循环，Llama、Granite 等模型均采用相同模式。该解释被接受，PR 最终获得批准。

- InternLM2Model.load_weights 手动实现与 AutoWeightsLoader 理念的差异 (design): gemini-code-assist[bot] 的建议未被采纳，作者的解释被接受。PR 最终获得批准。

风险与影响

- 风险：回归风险中等：改写了 InternLM2ForCausalLM 的 load_weights，若 AutoWeightsLoader 对 tie_word_embeddings 的 skip_prefixes 处理有误，可能导致权重加载失败。但该模式已在多个模型中验证，风险较低。另外，InternLM2Model 新增的 load_weights 为从外部复制的逻辑，已过 review。
- 影响：直接影响 InternLM2 系列模型的权重加载路径，使 InternLM2ForRewardModel 等复合模型能正确加载权重。对用户透明，无功能变化。团队可继续按此模式迁移其他模型至 AutoWeightsLoader，推动 Issue #15697 的完成。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #16383 previous attempt (reverted): PR 正文提及：先前尝试 (#16383) 因 deepseek_v2 的 KeyError 被回滚，但该 bug 与 internlm2 无关。
- PR #16453 revert of previous attempt: 回滚 #16383 的 PR，本 PR 确保不引入相同的问题。