PR #41694 完整报告

vllm-project/vllm

[DSV4] Add PP support for deepseek-v4

合并时间：2026-05-10 23:47

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41694

执行摘要

- 一句话：为 DeepSeek-V4 添加流水线并行支持
- 推荐动作：值得精读，展示了如何为复杂模型（含多流 hidden_states 和特殊注意力架构）添加 PP 支持，可作为其他模型 PP 集成的参考模板。

功能与动机

该 PR 是 DeepSeek V4 实现路线图（#40902）的一部分，目标是为 DeepSeek-V4 添加流水线并行支持，从而允许模型跨多个 GPU 部署以提升吞吐量。

实现拆解

1. 导入 PP 相关模块：在 *deepseek_v4.py* 中导入 get_pp_group、SupportsPP、PPMissingLayer、is_pp_missing_parameter。
2. 条件化初始化：在 DeepseekV4ForCausalLM 的 __init__方法中，根据 get_pp_group().is_first_rank/is_last_rank 决定是否初始化 embed_tokens、norm、_mtp_hidden_buffer、lm_head，非对应 rank 使用 PPMissingLayer 或 None。
3. 实现中间张量创建：新增 make_empty_intermediate_tensors 方法，返回形状为 (batch_size, hc_mult, hidden_size) 的 IntermediateTensors，该形状在 PP 阶段间传递多流 hidden_states。
4. 修改 forward 方法：根据 PP rank 决定是否执行 embedding（第一 rank）、norm+lm_head（最后 rank），中间 rank 直接传递 IntermediateTensors。
5. 更新权重加载：在 load_weights 中通过 is_pp_missing_parameter 跳过不在当前 rank 上的参数。
6. 更新文档：在 supported_models.md 中将 DeepSeek-V4 的 Pipeline Parallelism 列标记为支持。

关键文件：

- vllm/model_executor/models/deepseek_v4.py（模块 模型执行；类别 source；类型 core-logic；符号 make_empty_intermediate_tensors, DeepseekV4ForCausalLM）：核心实现文件，引入 PP 支持的所有逻辑：条件化初始化、中间张量创建、forward 修改、权重加载跳过。

- docs/models/supported_models.md（模块 文档；类别 docs；类型 documentation）：更新 DeepSeek-V4 的文档，标记 Pipeline Parallelism 支持。

关键符号：DeepseekV4ForCausalLM.init, DeepseekV4ForCausalLM.forward, make_empty_intermediate_tensors

关键源码片段

vllm/model_executor/models/deepseek_v4.py

核心实现文件，引入 PP 支持的所有逻辑：条件化初始化、中间张量创建、forward 修改、权重加载跳过。

```
# 导入 PP 相关模块
from vllm.distributed import get_pp_group
from vllm.model_executor.models.interfaces import SupportsPP
from .utils import PPMissingLayer, is_pp_missing_parameter

class DeepseekV4ForCausalLM(nn.Module, SupportsPP):
    def __init__(self, *, vllm_config: VllmConfig, prefix: str = ""):
        # ... ( 其他初始化 )

        # 根据 PP rank 条件化初始化组件
        if get_pp_group().is_first_rank:
            self.embed_tokens = VocabParallelEmbedding(
                config.vocab_size,
                config.hidden_size,
                quant_config=quant_config,
                prefix=f"{prefix}.embed_tokens",
            )
        else:
            self.embed_tokens = PPMissingLayer()

        if get_pp_group().is_last_rank:
            self.norm = RMSNorm(config.hidden_size, self.rms_norm_eps)
            self._mtp_hidden_buffer = torch.empty(
                vllm_config.scheduler_config.max_num_batched_tokens,
                self.hc_dim,
                dtype=vllm_config.model_config.dtype,
                device=self.device,
            )
        else:
            self.norm = PPMissingLayer()
            self._mtp_hidden_buffer = None

    def make_empty_intermediate_tensors(
        self,
        batch_size: int,
        dtype: torch.dtype,
        device: torch.device,
```

```
    ) -> IntermediateTensors:
        # PP 中间张量携带多流 hidden_states,
        # 形状为 (batch_size, hc_mult, hidden_size) ——
        # V4 在每个 token 进入第一个 decoder 层之前将嵌入扩展到 hc_mult 流,
        # 并在 hc_head() 折叠之前保持此形状。
        return IntermediateTensors({
            "hidden_states": torch.empty(
                batch_size,
                self.hc_mult,
                self.config.hidden_size,
                dtype=dtype,
                device=device,
            ),
        })
```

评论区精华

审阅者 jeejeelee 提出需要更新文档（https://docs.vllm.ai/en/latest/models/supported_models/#generative-models），该修改已在后续提交中完成并获 approve。

- 文档更新 (documentation): 已更新 supported_models.md

风险与影响

- 风险：风险包括：PP 配置错误导致初始化条件判断失败；中间张量形状与预期不匹配（尤其是 hc_mult 维度的处理）；MTP 缓冲区仅在最后 rank 分配，若其他 rank 误用可能引发空指针；与现有 cudagraph 和 aux_stream 的兼容性；缺少专门的 PP 测试覆盖。
- 影响：影响范围限于 DeepSeek-V4 模型，用户可通过设置 --pipeline-parallel-size >1 启用 PP。对于使用该模型的场景，PP 支持能显著提升模型吞吐量，降低单卡内存需求。团队需确保 PP 部署环境正确配置。
- 风险标记：核心路径变更，缺少测试覆盖，PP 配置风险

关联脉络

- PR #40902 [Roadmap] DeepSeek V4: 这是 PR 所关联的路线图 Issue，PR 是实现其中 Pipeline Parallelism 支持的部分。
- PR #40860 [Core] FP4 Indexer & MegaMoE initial support for DeepSeek-V4: 提供了 DeepSeek-V4 的基础模型支持，本 PR 在其基础上添加 PP。
- PR #40833 [MegaMoE] Continue work: 路线图中列出的后续工作，本 PR 的 PP 支持与之配合。