

PR #2267 完整报告

THUDM/slime

Fix model convert when use latest megatron

合并时间: 2026-08-16 17:45

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2267>

执行摘要

- 一句话: 修复最新 Megatron 下模型转换兼容性问题
- 推荐动作: 本 PR 值得精读, 因为它展示了如何通过小范围的兼容性修复来适配 Megatron 新版本, 特别是参数默认值、导入路径和构造签名的处理方式, 可作为后续适配其他模型插件时的参考。

功能与动机

PR 标题和描述指出, 使用最新版 Megatron 时模型转换工具存在兼容性问题, 需要调整以适配 Megatron 新版本的变化, 如参数命名空间、tokenizer 导入路径和注意力模块构造签名的新增 `name` 参数。

实现拆解

1. 更新 tokenizer 工具导入: 在 `slime/backends/megatron_utils/arguments.py` 中, 将 `_vocab_size_with_padding` 的导入修改为优先从 `megatron.core.tokenizers.utils.build_tokenizer` 导入, 并保留旧路径作为 fallback, 确保在不同 Megatron 版本下均能正常导入。
2. 设置 `enable_gloo_process_groups` 默认值: 在 `_set_default_megatron_args` 函数中, 当参数命名空间中不存在 `enable_gloo_process_groups` 时, 默认将其设置为 `True`, 以适配最新 Megatron 对进程组配置的要求。
3. 扩展注意层构造签名: 在 `slime_plugins/models/glm5/glm5.py`、`slime_plugins/models/qwen3_5.py` 和 `slime_plugins/models/qwen3_next.py` 中, 为各自的 `Attention` 类和 `DSAMLA_SelfAttention` 类的 `__init__` 方法添加可选的 `name: str | None = None` 参数, 并传递给父类, 以满足 Megatron 新版本构造时传入 `name` 的需求。
4. 转换工具新增参数: 在 `tools/convert_hf_to_torch_dist.py` 中, 为 `add_conversion_args` 添加 `--use-gated-attention` 命令行参数 (默认 `False`), 用于在转换时控制是否使用 gated attention, 并与已有的 `--attention-output-gate` 兼容。
5. 测试与验证: 本 PR 未包含直接测试文件改动; 需要依赖现有的测试框架进行兼容性验证。

关键文件:

- `slime/backends/megatron_utils/arguments.py` (模块 参数处理; 类别 `source`; 类型 `dependency-wiring`; 符号 `_set_default_megatron_args`): 核心参数处理文件, 涉及 tokenizer 导入路径和 `enable_gloo_process_groups` 默认值的设置, 直接影响 Megatron 兼容性。

- tools/convert_hf_to_torch_dist.py (模块 转换工具; 类别 source; 类型 core-logic; 符号 add_conversion_args) : 模型转换工具入口, 新增 --use-gated-attention 参数以支持 gated attention 的转换, 影响转换行为的可配置性。
- slime_plugins/models/qwen3_next.py (模块 模型插件; 类别 source; 类型 data-contract; 符号 Attention.init) : 为 Qwen3-Next 的注意层添加 name 参数, 适配 Megatron 新版本构造要求, 属于典型的接口兼容性修复。
- slime_plugins/models/qwen3_5.py (模块 模型插件; 类别 source; 类型 data-contract; 符号 Attention.init) : 与 Qwen3-Next 相同, 为 Qwen3.5 注意层添加 name 参数, 保证模型插件与最新 Megatron 兼容。
- slime_plugins/models/glm5/glm5.py (模块 模型插件; 类别 source; 类型 data-contract; 符号 DSAMLASelfAttention.init) : 为 GLM5 的 DSA 注意层添加 name 参数, 确保与最新 Megatron 兼容。

关键符号: _set_default_megatron_args, add_conversion_args, Attention.init, DSAMLASelfAttention.init

关键源码片段

slime/backends/megatron_utils/arguments.py

核心参数处理文件, 涉及 tokenizer 导入路径和 enable_gloo_process_groups 默认值的设置, 直接影响 Megatron 兼容性。

```
# slime/backends/megatron_utils/arguments.py 中的关键改动
from transformers import AutoConfig

try:
    # 优先从新的 Megatron tokenizer 工具模块导入
    from megatron.core.tokenizers.utils.build_tokenizer import vocab_size_with_padding as _vocab_size_with_padding
except ImportError:
    # 旧版 Megatron 回退到原路径
    from megatron.training.tokenizer.tokenizer import _vocab_size_with_padding

def _set_default_megatron_args(args):
    # ... 其他默认值设置 ...
    # 最新 Megatron 要求启用 gloo 进程组, 否则可能运行失败
    if not hasattr(args, "enable_gloo_process_groups"):
        args.enable_gloo_process_groups = True
    # ... 后续逻辑 ...
    return args
```

slime_plugins/models/qwen3_next.py

为 Qwen3-Next 的注意层添加 name 参数, 适配 Megatron 新版本构造要求, 属于典型的接口兼容性修复。

```
# slime_plugins/models/qwen3_next.py 中的构造方法
class Attention(HuggingfaceAttention):
```

```

def __init__(
    self,
    args,
    config,
    layer_number: int,
    cp_comm_type: str = "p2p",
    pg_collection=None,
    # 新增可选参数，用于接收 Megatron 传入的模块名称
    name: str | None = None,
):
    super().__init__(
        args,
        config,
        layer_number,
        cp_comm_type,
        pg_collection,
    )
    # ... 其余初始化逻辑 ...

```

slime_plugins/models/qwen3_5.py

与 Qwen3-Next 相同，为 Qwen3.5 注意层添加 `name` 参数，保证模型插件与最新 Megatron 兼容。

slime_plugins/models/qwen3_5.py 中的构造方法

```

class Attention(HuggingfaceAttention):
    def __init__(
        self,
        args,
        config,
        layer_number: int,
        cp_comm_type: str = "p2p",
        pg_collection=None,
        # 新增可选参数，用于接收 Megatron 传入的模块名称
        name: str | None = None,
    ):
        super().__init__(
            args,
            config,
            layer_number,
            cp_comm_type,
            pg_collection,
        )
        # ... 其余初始化逻辑 ...

```

评论区精华

本 PR 没有 review 评论或讨论线程，说明改动被直接接受，无公开争议。

- 暂无高价值评论线程

风险与影响

- 风险:

1. 回归风险: 对 `slime/backends/megatron_utils/arguments.py` 的导入路径调整和默认值设置可能影响其他依赖此工具的流程, 但已通过 `try/except` 保留旧路径, 风险较低。
2. 接口兼容性: 为 `attention` 模块增加 `name` 参数属于向后兼容的扩展, 但若其它代码直接调用这些类的构造而未传递该参数, 则不受影响; 但如果上游 `Megatron` 在构造时强制要求该参数, 则旧版本可能不兼容。
3. 参数默认值: `--use-gated-attention` 默认设置为 `False`, 可能改变已有转换行为, 需确认用户预期。
 - 影响: 影响范围主要集中在模型转换工具 (`tools/convert_hf_to_torch_dist.py`) 以及相关 `Megatron` 后端参数处理 (`slime/backends/megatron_utils/arguments.py`)。同时, 对多个模型插件 (`GLM5`、`Qwen3.5`、`Qwen3-Next`) 的注意层构造签名进行了扩展, 保证了与最新 `Megatron` 的兼容性。对现有用户来说, 转换工具新增参数不会破坏原有使用方式, 但可能需要调整命令行调用以利用新特性。对团队而言, 此改动降低了因 `Megatron` 版本升级带来的兼容性维护成本。
 - 风险标记: 核心路径变更 (参数处理), 缺少测试覆盖

关联脉络

- PR #2271 fix transform_ue8m0 in fp8 convert: 同属模型转换相关修复, 都涉及 `Megatron` 到 `HF` 的转换流程兼容性。