

PR #2093 完整报告

THUDM/slime

Support GLM-5.2

合并时间: 2026-06-17 07:04

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2093>

执行摘要

- 一句话: 新增 GLM-5.2 744B-A40B 模型训练支持
- 推荐动作: 该 PR 是典型的模型接入变更, 核心逻辑改动集中在路由和配置, 适合作为 DSA 系列模型接入的参考示例。建议阅读 `slime/backends/megatron_utils/megatron_to_hf/__init__.py` 的分支扩展方式和 `slime_plugins/models/glm5/glm5.py` 中的 pipeline 分割校验逻辑。

功能与动机

支持开源 GLM-5.2 模型 (744B-A40B) 的训练, 该模型使用 DSA 跨层索引共享架构, 需要适配 slime 的 Megatron 训练与 SGLang rollout 流程。PR body 未提供额外细节, 但从变更内容推断是用户 / 团队需求。

实现拆解

1. 新增训练入口脚本: 添加 `scripts/run-glm5.2-744B-A40B.sh`, 包含完整的 Ray 集群启动、SGLang 服务部署、Megatron 训练参数 (TP=4, PP=8, CP=8, EP=32) 以及 GRPO 优化器配置。脚本从 `scripts/models/glm5.2-744B-A40B.sh` 加载模型结构参数。
2. 新增模型参数配置: `scripts/models/glm5.2-744B-A40B.sh` 定义 GLM-5.2 的 MoE、注意力、隐藏层等参数, 并配置 `--allgather-cp` 标志以启用 DSA 在 context parallel 下的 allgather 布局。
3. 更新模型 spec provider: 在 `slime_plugins/models/glm5/glm5.py` 中调整注释和模型名称引用, 反映 GLM-5.2 的新命名 (GLM-5.2 744B-A40B)。
4. 扩展权重转换路由: 在 `slime/backends/megatron_utils/megatron_to_hf/__init__.py` 的 `_convert_to_hf_core` 函数中增加 `glmmoedsa` 分支, 使其使用与 `deepseekv3` 相同的转换逻辑。
5. 注册 bridge 别名: 在 `slime_plugins/mbridge/deepseek_v32.py` 中为 `DeepseekV32Bridge` 添加 `glm_moe_dsa` 注册名, 使 GLM-5.2 的 HF config (`model_type: glm_moe_dsa`) 能被映射到正确桥接器。
6. 添加转换工具参数: 在 `tools/convert_hf_to_torch_dist.py` 的 `add_conversion_args` 中注册 `--allgather-cp` 参数 (转换时为 `no-op`), 避免脚本因未知参数报错。
7. 为现有脚本添加 `--allgather-cp`: 修改 `scripts/run-deepseek-r1.sh`、`scripts/run-glm4.7-355B-A32B.sh`、`scripts/run-qwen3-next-80B-A3B.sh`、

scripts/low_precision/run-kimi-k2-Thinking-int4.sh, 统一增加该参数确保一致性。

8. 编写中英文文档: 添加 docs/en/examples/glm5.2-744B-A40B.md 和 docs/zh/examples/glm5.2-744B-A40B.md, 详细说明环境准备、checkpoint 转换、训练执行步骤及参数含义。

关键文件:

- scripts/run-glm5.2-744B-A40B.sh (模块 启动脚本; 类别 other; 类型 core-logic) : 新增的 GLM-5.2 训练主入口, 整合了 Ray 集群、SGLang 部署、Megatron 训练参数 (TP/PP/CP/EP) 及 GRPO 优化器设置, 是用户启动训练的起点。
- scripts/models/glm5.2-744B-A40B.sh (模块 模型配置; 类别 other; 类型 data-contract) : 定义了 GLM-5.2 的模型结构参数 (层数、隐藏维度、专家数等), 通过 MODEL_ARGS 传递给 Megatron, 是训练配置的核心。
- slime_plugins/models/glm5/glm5.py (模块 注意力模块; 类别 source; 类型 data-contract; 符号 get_glm5_spec) : 提供 GLM-5 系列的模型 spec provider (get_glm5_spec), 本次更新了注释以准确反映 GLM-5.2 的命名和参数, 其非共享路径逻辑无变化。
- slime/backends/megatron_utils/megatron_to_hf/__init__.py (模块 转换路由; 类别 source; 类型 core-logic; 符号 _convert_to_hf_core) : 权重转换路由函数 _convert_to_hf_core 中新增了 'glmmodsa' 模型名识别, 将其导向已有的 deepseekv3 转换逻辑, 是支持 GLM-5.2 权重导出的关键改动。
- slime_plugins/mbridge/deepseek_v32.py (模块 桥接模块; 类别 source; 类型 core-logic; 符号 DeepseekV32Bridge) : 为 DeepseekV32Bridge 增加 'glm_moe_dsa' 注册名, 使 GLM-5.2 的 HF 配置能被自动映射到正确的桥接器, 是模型加载的关键环节。
- 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) : 添加了 --allgather-cp 参数注册, 确保 GLM-5.2 的转换脚本不会因该自定义参数报错 (虽然转换时 no-op)。
- docs/en/examples/glm5.2-744B-A40B.md (模块 英文文档; 类别 docs; 类型 documentation) : 英文版训练文档, 包含环境搭建、checkpoint 转换、训练执行等完整步骤, 是用户上手的主要参考资料。
- docs/zh/examples/glm5.2-744B-A40B.md (模块 中文文档; 类别 docs; 类型 documentation) : 中文版训练文档, 内容与英文版一致, 方便中文用户使用。

关键符号: get_glm5_spec, _convert_to_hf_core, add_conversion_args, DeepseekV32Bridge

关键源码片段

slime_plugins/models/glm5/glm5.py

提供 GLM-5 系列的模型 spec provider (get_glm5_spec), 本次更新了注释以准确反映 GLM-5.2 的命名和参数, 其非共享路径逻辑无变化。

```
# slime_plugins/models/glm5/glm5.py (head) — get_glm5_spec 函数片段
```

```
# Optional cross-layer index-sharing schedule. Present on DSA checkpoints that
```

```
# only store indexer weights on a subset of "computing" layers (e.g. GLM-5.2
# 744B-A40B). When absent, every layer computes its own top-k (plain
# DSA) and ``DSAMLASelfAttention`` runs the non-shared path.
config.index_topk_freq = getattr(hf_config, "index_topk_freq", 1) or 1
config.index_skip_topk_offset = getattr(hf_config, "index_skip_topk_offset", 0) or 0
```

... (后续 pipeline 分割校验)

slime/backends/megatron_utils/megatron_to_hf/__init__.py

权重转换路由函数 `_convert_to_hf_core` 中新增了 'glmmoedsa' 模型名识别，将其导向已有的 deepseekv3 转换逻辑，是支持 GLM-5.2 权重导出的关键改动。

slime/backends/megatron_utils/megatron_to_hf/__init__.py (head) — `_convert_to_hf_core` 分支

```
def _convert_to_hf_core(args, model_name, name, param):
    if "minimaxm2" in model_name or "minimax_m2" in model_name:
        converted_named_tensors = convert_minimax_m2_to_hf(args, name, param)
    elif "glm4moelite" in model_name or "deepseekv3" in model_name or "glmmoedsa" in model_
    name:
        # GLM-5.2 的 model_type 为 "glm_moe_dsa", 指向此分支
        converted_named_tensors = convert_deepseekv3_to_hf(args, name, param)
    # ... 其余分支
```

评论区精华

该 PR 无任何 review 评论或讨论，可能为内部直接合并。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 缺少测试覆盖：未包含任何新测试，GLM-5.2 的模型 spec、权重转换、bridge 路由等核心路径无自动化验证，存在回归风险。
2. 跨层索引共享的 pipeline 限制：`get_glm5_spec` 函数中包含了将 pipeline 分割必须避开 skip layer 的断言 (`AssertionError`)，若用户自行调整 `--pipeline-model-parallel-size` 而不匹配 `decoder-first-pipeline-num-layers` 和 `decoder-last-pipeline-num-layers`，训练会在启动时崩溃。文档虽给出了推荐值，但未提供自动校验。
3. `allgather-cp` 参数扩散：在多个已有运行脚本中追加 `--allgather-cp`，若这些模型实际并不需要该参数（如非 DSA 模型），可能会导致未预期的行为或性能退化。当前默认值未在参数定义中明确，需确认其他模型是否兼容。
4. bridge 映射耦合：将 `glm_moe_dsa` 直接映射到 `DeepseekV32Bridge`，未来若 GLM-5.2 的权重布局发生变化，需额外维护路径。
- 影响：对用户：新用户可参照新增文档训练 GLM-5.2 744B-A40B 模型，现有其他模型的运行脚本也因 `--allgather-cp` 的加入而改变（兼容性风险见上）。对系统：扩展了 `megatron_to_hf` 的模型类型分支，增加了 `mbridge` 的注册条目，变更范围较小且向后兼容（新增分支不影响既有路径）。对

团队：需要对 GLM-5.2 的长期维护负责，尤其在权重转换和 bridge 层面。 - 风险标记：缺少测试覆盖，跨层索引共享管道分割限制，allgather-cp 参数扩散风险，bridge 映射硬编码

关联脉络

- PR #2072 [docker] upgrade sglang to v0.5.13: 该 PR 为 sglang 引入了 GLM-5 的跨层索引共享支持，是 GLM-5.2 训练的基础依赖。
- PR #2082 Overlapping data loading and sglang initialization: 改进了 sglang 引擎的启动流程，与本 PR 的训练脚本中 SGLang 服务管理相关。