

PR #2660 完整报告

radixark/miles

[AMD] Point the ROCm image at the v0.5.16 wheels release with TE 2.17.0

合并时间: 2026-08-20 10:52

原文链接: <http://prhub.com.cn/radixark/miles/pull/2660>

执行摘要

- 一句话: ROCm 镜像升级 wheels 并切换 FP8 块缩放策略
- 推荐动作: 值得轻度关注, 尤其是 AMD 团队。建议精读以理解 TE 2.17.0 的行为变化, 并测试 FP8 缩放调整对模型收敛的影响。由于改动小, 没有复杂的架构决策。

功能与动机

PR 标题和提交信息表明需要将 ROCm 镜像指向 v0.5.16 wheels 发布 (包含 TE 2.17.0), 同时调整 FP8 缩放设置。由于 TE 版本升级, NVTE_FP8_BLOCK_SCALING_FP32_SCALES 的值从 '1' 改为 '0', 可能因为 TE 2.17.0 不再推荐 FP32 缩放, 或需要与 wheels 中的新构建保持一致。

实现拆解

1. 修改 docker/Dockerfile.rocm, 将 WHEELS_TAG_ROCM 从 'rocm720-gfx950-v0.5.14' 更新为 'rocm720-gfx950-v0.5.16', 确保镜像使用包含 TE 2.17.0 的预编译 wheels。
2. 修改 scripts/amd/run_deepseek_v4.py, 在启用 FP8 训练时, 将 NVTE_FP8_BLOCK_SCALING_FP32_SCALES 从 '1' 改为 '0', 并移除对应的注释。这意味着块缩放使用非 FP32 尺度 (可能为 BF16), 与 TE 2.17.0 的默认行为对齐。
3. 测试配套: 无直接测试文件变更, 但需要后续的镜像构建和训练运行验证。

关键文件:

- scripts/amd/run_deepseek_v4.py (模块 AMD 脚本; 类别 source; 类型 core-logic; 符号 _train): 修改 FP8 训练时的缩放环境变量, 直接影响数值稳定性和性能
- docker/Dockerfile.rocm (模块 Docker; 类别 infra; 类型 infrastructure): 更新 wheels 标签, 确保镜像使用包含 TE 2.17.0 的预编译包

关键符号: _train

关键源码片段

`scripts/amd/run_deepseek_v4.py`

修改 FP8 训练时的缩放环境变量, 直接影响数值稳定性和性能

```
if args.fp8_training:
    misc_args += "--transformer-impl transformer_engine " "--bf16 " "--fp8-format e4m3 " "--fp8-
```

```
recipe blockwise "  
# TE 2.17.0 更新后, gfx950 改用非 FP32 块缩放, 这里关闭 FP32 尺度  
misc_args += "'--train-env-vars '{\"NVTE_FP8_BLOCK_SCALING_FP32_SCALES\":\"0\"}' '"  
# ROCm TE MoE FP8 缺乏融合的 wgrad 累加, 禁用该融合  
misc_args += "--no-gradient-accumulation-fusion "
```

评论区精华

无人工 review 评论, claude[bot] 自动审查因 fork 而禁用, 仅 guapisolo 批准。

- 暂无高价值评论线程

风险与影响

- 风险: 风险较低。wheels 版本升级可能引入 TE 2.17.0 的其他行为变化, 影响训练稳定性和性能。FP8 缩放策略从 FP32 改为非 FP32, 可能影响模型精度和收敛, 需要在 gfx950 上实际训练验证。若不匹配, 可能导致数值不稳定或性能下降。
- 影响: 影响范围限于 ROCm 后端 (AMD GPU) 的 DeepSeek-V4 训练。更新 Dockerfile 会影响所有从该镜像构建的部署。FP8 缩放改变会影响训练精度和速度, 但对用户透明。团队需重新构建镜像并验证训练任务。
- 风险标记: 依赖版本升级, 数值精度变化, 缺少测试覆盖

关联脉络

- PR #2600 fix(docker): pin cutlass-dsl 4.6.2 and flashinfer 0.6.15.post1 over the sglang base: 相同领域: docker 和 wheels 版本控制