PR #22971 完整报告

sgl-project/sglang

[AMD][diffusion] Temporal-unfolded batched Conv2D for ROCm VAE decode

合并时间：2026-05-08 17:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22971

执行摘要

- 一句话：ROCm VAE 解码：时间展开 Conv2D 加速 3.6%
- 推荐动作：值得阅读学习时间展开的实现技巧和平台抽象层的优化模式。建议后续跟进未采纳的评审建议，增强替换代码的鲁棒性（如声明支持的卷积参数范围）。

功能与动机

ROCm 平台上 3D 卷积效率低于 2D 卷积。本 PR 利用时间展开技巧，将 Conv3d 等价转换为 batched Conv2D，利用更高效的 2D 卷积实现加速，同时提供 BF16 选项。

实现拆解

实现可分为以下步骤：

1. 环境变量注册：在 *envs.py* 的 EnvironmentVariables 类中新增 SGLANG_USE_ROCM_VAE_CONV2D 和 SGLANG_USE_ROCM_VAE_CONV2D_BF16 两个布尔类型变量，并在 *environment_variables* 字典中注册，供运行时查询。
2. 核心算法：在 rocm.py 中实现静态方法 _conv3d_as_batched_conv2d。该方法对输入张量沿时间轴进行 unfold，将 (N, C, T, H, W) 形状转换为 (NT', KtC, H, W) 形状，并调用 F.conv2d 计算，最后将结果折叠回原有时间维度。当启用 BF16 时，先将输入转换为 bf16，计算完毕再转回原精度。
3. 模块替换：实现静态方法 _replace_conv3d_with_conv2d，递归遍历 VAE 模型，找到所有 nn.Conv3d 实例（通常为 CausalConv3d，假设 groups=1, dilation=1）。在替换时预先将 3D 权重转换为 2D 权重并缓存为 weight_2d 属性，同时替换 forward 为 _patched_forward，后者调用 _conv3d_as_batched_conv2d。
4. 集成到优化流程：在 RocmPlatform.optimize_vae 中，在原有的 GroupNorm 替换之后，检查 SGLANG_USE_ROCM_VAE_CONV2D 或 SGLANG_USE_ROCM_VAE_CONV2D_BF16 环境变量，若为真则执行 Conv3d 替换，并记录替换数量。
5. 精度与性能验证：PR 提供了 Wan2.2 T2V 模型上的精度对比（PSNR 38.12dB）和端到端加速（3.6%），确保替换前后输出视觉一致。

关键文件：

- python/sglang/multimodal_gen/runtime/platforms/rocm.py（模块 ROCm 平台；类别 source；类型 core-logic；符号 _conv3d_as_batched_conv2d,

_replace_conv3d_with_conv2d, _patched_forward）：核心实现文件，包含时间展开 batched Conv2D 替换算法和模块替换函数

- python/sglang/multimodal_gen/envs.py（模块 环境变量；类别 source；类型 configuration）：定义两个新的环境变量控制 Conv2D 替换行为

关键符号：_conv3d_as_batched_conv2d, _replace_conv3d_with_conv2d, _patched_forward

评论区精华

评审机器人 gemini-code-assist[bot] 在三个层面提出改进建议：签名泛化（使用 *args, **kwargs）、参数完整性（传递 groups 和 dilation）、性能优化（缓存 weight 转换）。其中权重缓存在最终代码中已实现（替换时预计算 weight_2d 并存储在模块属性中），其余两项未采纳。最终由 HaiShaw 批准合并。

- Monkey-patched forward 签名健壮性 (correctness): 未采纳，PR 保持简单签名
- F.conv2d 未传递 groups 和 dilation 参数 (correctness): 未采纳，假设默认值
- Weight 转换应缓存 (performance): 已在替换时预计算 weight_2d 并缓存，评论已过时

风险与影响

- 风险：主要风险集中在泛化性：假设所有 Conv3d 的 groups=1 且 dilation=1，若未来 VAE 变体使用非默认参数，替换将产生错误结果。此外，weight 转换在模块替换时只计算一次，但 unfold 操作在每次 forward 都会执行，可能对短序列增加小幅开销。替换逻辑未包含参数校验，不兼容情况会静默失败。
- 影响：影响范围小，仅对 ROCm 后端的扩散模型 VAE 解码路径有效，且需显式设置环境变量。对 CUDA 或其他后端无影响。用户启用后，Wan2.2 等视频模型可获得 3%-4% 的端到端加速，且输出质量几乎无损。代码改动量小，易于维护。
- 风险标记：未处理 groups/dilation 参数，forward 签名假设，不支持非标准 Conv3d

关联脉络

- 暂无明显关联 PR