

PR #27884 完整报告

sgl-project/sglang

Fix deep seek ocr2 image processing

合并时间: 2026-06-17 12:59

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27884>

执行摘要

- 一句话: 修复 DeepSeek-VL2 基准测试图片处理
- 推荐动作: 值得合并: 修复明确且无副作用。建议后续为该函数补充单元测试 (如使用 mock processor 测试各分支), 提升可维护性。

功能与动机

DeepSeek VL2 的 `format_messages_v2` 要求通过 `conversations=` 传入包含 `<image>` 标记的字符串, 而原有的通用处理器分支无法处理该格式, 导致基准测试数据创建失败或 token 计算错误。PR body 明确指出了此兼容性问题。

实现拆解

修改 `python/sglang/benchmark/datasets/image.py` 中的 `create_mm_data_row` 函数, 在该函数已有的 `KimiK25Processor` 分支之后新增一个 `elif` 分支, 专门处理 `DeepseekVLV2Processor` 类型:

1. 使用 `processor(conversations=prompt_str, images=images, inference_mode=True)` 调用, 其中 `conversations` 是 DeepSeek-VL2 期望的包含 `<image>` 占位符的字符串。
2. 从返回结果的 `input_ids` 字段获取 `prompt_len`。
3. 其余逻辑 (`text-only` 计算、`vision token` 差值) 保持不变。

关键文件:

- `python/sglang/benchmark/datasets/image.py` (模块 基准测试; 类别 `source`; 类型 `core-logic`): `benchmark` 图片数据集行创建函数, 新增 `DeepseekVLV2Processor` 分支以支持特殊的 `conversations` 参数格式。

关键符号: `create_mm_data_row`

关键源码片段

`python/sglang/benchmark/datasets/image.py`

`benchmark` 图片数据集行创建函数, 新增 `DeepseekVLV2Processor` 分支以支持特殊的 `conversations` 参数格式。

```
# 在计算 total tokens 的 if-elif-else 链中新增 DeepseekVLV2Processor 分支
if type(processor).__name__ == "KimiK25Processor":
```

```

medias = [{"type": "image", "image": img} for img in images]
prompt_len = processor(
    text=prompt_str,
    medias=medias,
    return_tensors="pt",
)["input_ids"].numel()
elif type(processor).__name__ == "DeepseekVLV2Processor":
    # DeepSeek-VL2 使用 conversations 格式而非 text+images 列表
    # 需传入带有 <image> 占位符的 prompt_str
    result = processor(
        conversations=prompt_str, # 期望格式如 "<image>Describe..."
        images=images,
        inference_mode=True, # 启用推理模式以正确编码
    )
    prompt_len = result.input_ids.numel()
else:
    # 通用分支：适用于 InternVL、Qwen 等处理器
    prompt_len = processor(
        text=[prompt_str],
        images=images,
        padding=False,
        return_tensors="pt",
    )["input_ids"].numel()

```

评论区精华

该 PR 未产生 review 讨论（review comments 为 0），仅由 mingfeima 直接 approving 合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅新增一个类型分支，不影响其他处理器流程；代码逻辑简单清晰。但未添加对应单元测试，未来若有类似格式变更可能被遗漏。
- 影响：影响范围局限于 sglang benchmark 工具中 DeepSeek-VL2/OCR2 模型的数据行创建场景。对生产推理无影响，仅影响基准测试的正确性。
- 风险标记：缺少测试覆盖

关联脉络

- PR #21472 Fix PicklingError with --backend diffusers on non-T2I models: 同一仓库中的 diffusion/bugfix PR，均涉及 benchmark 或 processor 兼容性修复，且同属 multimodal 领域。