

PR #6389 完整报告

verl-project/verl

[perf] feat: Add Qwen3.5 MFU flops estimation

合并时间: 2026-07-07 17:08

原文链接: <http://prhub.com.cn/verl-project/verl/pull/6389>

执行摘要

- 一句话: 添加 Qwen3.5 系列 MFU/FLOPs 估算支持
- 推荐动作: 建议阅读此 PR 以了解 Qwen3.5 混合注意力的 FLOPs 估算方法, 特别是 `_count_qwen3_5_layer_types` 中根据 `layer_types` 或回退公式识别层类型的设计, 以及 GatedDeltaNet 循环 FLOPs 的计算方式。review 中对因果注意力因子的修正也值得关注, 类似问题可能在其它模型估算中复现。

功能与动机

PR 描述提到需要为 `qwen3_5` 和 `qwen3_5_moe` 添加 MFU/FLOPs 估算, 以支持性能监控。Qwen3.5 引入了混合注意力 (Hybrid Attention), 包括 full attention 和 GatedDeltaNet linear attention, 需要专门的 FLOPs 计算公式。同时需要计入 ViT 部分的 FLOPs。

实现拆解

1. 在 `verl/utils/flops_counter.py` 中新增四个私有函数:
 - `_count_qwen3_5_layer_types`: 从模型配置读取 `layer_types` 或按 `full_attention_interval` 回退, 统计 full attention 层数和 linear attention 层数。
 - `_compute_qwen3_5_hybrid_attn_params`: 基于配置计算 full attention (含 2x q_proj 输出门)、linear attention (含卷积投影) 的参数量, 并汇总所有注意力层的总线性参数。
 - `_compute_qwen3_5_gdn_recurrence_flops`: 计算 GatedDeltaNet 循环部分的 FLOPs (维度乘积乘以 token 数和 linear 层数)。
 - `_estimate_qwen3_5_flops`: 主估算函数, 整合文本 MLP (区分 Dense/MoE)、注意力参数、embedding+LM head、全注意力二次项、GatedDeltaNet 循环, 以及 ViT 部分 (复用 `_estimate_qwen3_vit_flops`)。
2. 在 `ESTIMATE_FUNC` 字典中注册 `qwen3_5` 和 `qwen3_5_moe` 到 `_estimate_qwen3_5_flops`。
3. 修改 `tests/utils/test_flops_counter.py`:
 - 在 `VALID_CONFIG_TYPE` 中添加 `qwen3_5` 和 `qwen3_5_moe`。
 - 新增两个测试配置, 分别对应 Qwen3.5-27B (纯 Dense) 和 Qwen3.5-35B-A3B (MoE), 提供完整的 `text_config`、`vision_config`、`batch_seqlens_tuple`、`images_seqlens_tuple` 和手动计算验证过的 `expected_flops_tuple`。

关键文件：

- verl/utils/flops_counter.py（模块 FLOPs 估算；类别 source；类型 core-logic；符号 `_count_qwen3_5_layer_types`, `_compute_qwen3_5_hybrid_attn_params`, `_compute_qwen3_5_gdn_recurrence_flops`, `_estimate_qwen3_5_flops`）：核心实现文件，新增四个函数注册到 `ESTIMATE_FUNC` 字典，为 Qwen3.5 系列提供 FLOPs 估算
- tests/utils/test_flops_counter.py（模块 FLOPs 测试；类别 test；类型 test-coverage）：测试配套，新增两个回归测试用例覆盖 Dense 和 MoE 变体，验证估算值与手动计算一致

关键符号：`_count_qwen3_5_layer_types`, `_compute_qwen3_5_hybrid_attn_params`, `_compute_qwen3_5_gdn_recurrence_flops`, `_estimate_qwen3_5_flops`

关键源码片段

tests/utils/test_flops_counter.py

测试配套，新增两个回归测试用例覆盖 Dense 和 MoE 变体，验证估算值与手动计算一致

```
# tests/utils/test_flops_counter.py
```

```
VALID_CONFIG_TYPE = {  
    "llama",  
    "qwen2",  
    "qwen3",  
    "qwen3_moe",  
    "qwen3_5", # 新增  
    "qwen3_5_moe", # 新增  
    "deepseek_v3",  
    "mistral",  
    "gemma3_text",  
    "apertus",  
}
```

```
# 在 CONFIG 字典中新增 qwen3_5 和 qwen3_5_moe 条目
```

```
# qwen3_5 配置示例：
```

```
"qwen3_5": {  
    "config": { # Qwen/Qwen3.5-27B  
        "model_type": "qwen3_5",  
        "text_config": {  
            "vocab_size": 248320,  
            "hidden_size": 4096,  
            "intermediate_size": 12288,  
            "num_hidden_layers": 32,  
            "num_attention_heads": 16,  
            "num_key_value_heads": 4,  
            "head_dim": 256,  
            "linear_conv_kernel_dim": 4,  
            "linear_key_head_dim": 128,  
            "linear_value_head_dim": 128,  
            "linear_num_key_heads": 16,
```

```

        "linear_num_value_heads": 32,
        "layer_types": ["linear_attention" if bool((i + 1) % 4) else "full_attention" for i in
            range(32)],
    },
    "vision_config": {
        # ... 标准 Qwen3 视觉配置
    },
},
"batch_seqlens_tuple": ([512, 1024, 2048], [4096, 4096, 4096]),
"images_seqlens_tuple": ([512, 1024, 2048], [4096, 4096, 4096]),
"expected_flops_tuple": (
    206090394402816 / 1e12,
    724521757704192 / 1e12,
),
},

```

评论区精华

review 过程中，gemini-code-assist[bot] 提出了两个高优先级问题：

1. 在 `_estimate_qwen3_5_flops` 中缺失 `embedding` 参数的 FLOPs 计算（只计了 LM head），建议改为 $2 * \text{vocab_size} * \text{hidden_size}$ 以与其他模型一致。
 2. full attention 的因果掩码应使二次项因子为 6 而非 12，否则会高估 2 倍。作者 liangxuZhang 均回复 "fixed"，确认已修正。最终得到 wuxibin89 的 approval。
- full attention 因果因子应为 6 而非 12 (correctness): 作者 liangxuZhang 回复 "fixed"，已修正代码中的因子。
 - embedding 参数应计入 dense FLOPs (correctness): 作者 liangxuZhang 回复 "fixed"，已修正为 $\text{emd_and_lm_head_N} = 2 * \text{vocab_size} * \text{hidden_size}$ 。

风险与影响

- 风险：主要风险在于 FLOPs 估算公式的准确性：如果 full attention 的因果因子或 linear attention 的循环公式存在误差，会导致 MFU 计算偏差。已在 review 中修正因果因子，但 GatedDeltaNet 循环的系数（15）是否适用于所有配置需要持续验证。此外，新增代码仅影响 FLOPs 估算，不影响训练核心路径，回归风险低。测试配置基于特定模型设置，若 Qwen3.5 后续版本更改层类型配置格式，估算可能失效。
- 影响：对用户：使用 Qwen3.5 和 Qwen3.5-MoE 模型的用户现在可以在训练中获得准确的 MFU 指标，便于性能调优。对系统：在 `flops_counter.py` 中新增约 100 行代码，不影响现有模型估算。对团队：提供了 Qwen3.5 混合注意力的 FLOPs 计算参考，便于后续模型扩展。影响范围仅限于 FLOPs 计数模块，影响程度中等。
- 风险标记：估算偏差风险，模型配置依赖

关联脉络

- 暂无明显关联 PR