PR #30275 完整报告

sgl-project/sglang

[Model] Support LongCat 2.0 FP8

合并时间：2026-07-07 19:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30275

执行摘要

- 一句话：支持 LongCat 2.0 FP8 模型在 8x B300 上部署与推理
- 推荐动作：本 PR 的 ngram token table 跳过掩码设计和 DSA indexer head padding 技巧值得精读；建议后续补充端到端集成测试并监控 decode 性能。

功能与动机

支持 meituan-longcat/LongCat-2.0-FP8 在 SGLang 中运行。原 PR #30202 提供了基础实现，但在实际 8x B300 服务验证中发现精度问题，需修复后合入。

实现拆解

1. LongCat 配置自动发现：在 python/sglang/srt/utils/hf_transformers/config.py 中新增 _try_load_longcat_config 函数，通过检查 architectures 字段是否包含 LongCat 架构名，自动加载自定义配置类 longcat_flash.py，并处理 model_type 为 null 的兼容问题。
2. 模型层调整：修改 python/sglang/srt/models/longcat_flash.py，在 LongcatFlashDecoderLayer.forward 中传递 prev_topk_indices 参数以支持 DSA indexer 跨层传播；LongcatFlashModel.forward 同样传递 topk_indices。
3. 统一 ngram token 表更新：新增 python/sglang/srt/model_executor/ngram_token_table.py，提供 update_ngram_token_table_after_sampling 函数，根据 skip_token_table_update 掩码跳过未完成的 chunked prefill 请求的伪 token，防止污染 token 表。对应测试文件覆盖三个场景。
4. 重构 ngram 嵌入层：修改 python/sglang/srt/layers/n_gram_embedding.py，移除 decode 分支，统一使用 compute_n_gram_ids；新增 eos_token_id 参数；将 VocabParallelEmbedding 的 enable_tp 改为 use_attn_tp_group，确保与 LongCat 的 TP 设置一致。
5. 增强 DSA Indexer：在 python/sglang/srt/layers/attention/dsa/dsa_indexer.py 中添加 _pad_heads_for_deep_gemm 静态方法（当 head 数 < 32 时填充）和 _mask_init_and_local_tokens 方法（遮蔽初始 / 局部 token 的 logits），支持 LongCat 的 ngram 预测策略。
6. 填充 ngram 元数据：在 scheduler.py、forward_batch_info.py、base_runner.py 中为 warmup 和 CUDA graph dummy batch 填充 ngram_embedding_info 的必要字段，避免图捕获阶段报错。

关键文件：

- test/registered/unit/model_executor/test_ngram_token_table.py（模块 ngram 测试；类别 test；类型 test-coverage；符号 _make_ngram_info, TestNgramTokenTableUpdate, test_chunked_prefill_mask_skips_pseudo_next_token, test_all_requests_masked_does_not_update_table）：测试覆盖 ngram token 表更新的三个核心场景，确保跳过掩码逻辑正确。
- python/sglang/srt/model_executor/ngram_token_table.py（模块 ngram 表；类别 source；类型 data-contract；符号 update_ngram_token_table_after_sampling）：核心新文件，实现 ngram token 表更新逻辑，包含 chunked prefill 跳过掩码机制。
- python/sglang/srt/layers/attention/dsa/dsa_indexer.py（模块 注意力索引；类别 source；类型 core-logic；符号 _pad_heads_for_deep_gemm, _mask_init_and_local_tokens）：修改 DSA indexer 以支持 LongCat 的 head padding 和 init/local token 掩码。
- python/sglang/srt/utils/hf_transformers/config.py（模块 配置解析；类别 source；类型 core-logic；符号 _try_load_longcat_config）：实现 LongCat 配置自动发现，解决 model_type 为 null 时的加载问题。
- python/sglang/srt/layers/n_gram_embedding.py（模块 ngram 嵌入；类别 source；类型 dependency-wiring）：重构 ngram 嵌入层，移除 decode 分支、统一 compute_n_gram_ids、调整 TP 参数。
- python/sglang/srt/models/longcat_flash.py（模块 模型层；类别 source；类型 data-contract）：LongCat 模型定义，修改 forward 传递 topk_indices 以支持 DSA indexer。
- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic）：为 warmup 和 CUDA graph dummy batch 设置 ngram embedding 元数据。
- python/sglang/srt/model_executor/forward_batch_info.py（模块 前向批次；类别 source；类型 data-contract）：在 ForwardBatch 中添加 ngram 相关字段支持。

关键符号：update_ngram_token_table_after_sampling, _try_load_longcat_config, _pad_heads_for_deep_gemm, _mask_init_and_local_tokens, NgramEmbedding.init, LongcatFlashDecoderLayer.forward, LongcatFlashModel.forward

关键源码片段

python/sglang/srt/model_executor/ngram_token_table.py

核心新文件，实现 ngram token 表更新逻辑，包含 chunked prefill 跳过掩码机制。

```
"""Utilities for updating LongCat ngram embedding token tables."""

from __future__ import annotations

import torch

from sglang.jit_kernel.ngram_embedding import update_token_table

def update_ngram_token_table_after_sampling(
    *,
```

```
    ngram_embedding_info,
    next_token_ids: torch.Tensor,
    req_pool_indices: torch.Tensor,
    seq_lens: torch.Tensor,
    batch_size: int,
) -> bool:
    """Update the ngram token table with sampled tokens.

    Returns whether the token table was updated.
    """
    skip_token_table_update = ngram_embedding_info.skip_token_table_update
    if skip_token_table_update is not None:
        # 跳过 chunked prefill 未完成的请求：它们的采样 token 是伪预测，不能污染 token 表
        indices = (~skip_token_table_update).nonzero(as_tuple=True)[0]
        if indices.numel() == 0:
            return False
        update_token_table(
            ne_token_table=ngram_embedding_info.token_table,
            tokens=next_token_ids[indices].to(torch.int32),
            row_indices=req_pool_indices[indices],
            column_starts=seq_lens[indices].to(torch.int32),
            req_lens=torch.ones(
                indices.numel(), dtype=torch.int32, device=next_token_ids.device
            ),
            ignore_tokens=None,
        )
        return True

    # 无 mask 时，直接写入所有 token
    ngram_embedding_info.out_column_starts[:batch_size] = seq_lens
    ngram_embedding_info.out_req_lens[:batch_size] = 1
    update_token_table(
        ne_token_table=ngram_embedding_info.token_table,
        tokens=next_token_ids.to(torch.int32),
        row_indices=req_pool_indices,
        column_starts=ngram_embedding_info.out_column_starts,
        req_lens=ngram_embedding_info.out_req_lens,
        ignore_tokens=None,
    )
    return True
```

评论区精华

该 PR 由 ispobock 快速审批通过，无实质性设计讨论。内部协作体现在 commit log 中，BBuf 和 sunjiaqi11 合作修复了 serving 精度问题。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 配置依赖风险：_try_load_longcat_config 依赖于 HuggingFace 配置的 architectures 字段，若上游模型变更可能导致加载失败。
  - token 表污染风险：skip_token_table_update 掩码的正确性依赖上游正确设置，若其他模型误用可能导致 ngram 表被伪 token 污染。
  - DSA indexer 逻辑耦合：_pad_heads_for_deep_gemm 和 _mask_init_and_local_tokens 仅适用于 LongCat，若被其他模型误用可能产生错误结果。
  - decode 路径合并风险：统一使用 compute_n_gram_ids 替代原有的 decode 分支，虽然简化了逻辑，但可能对 decode 阶段的性能有隐忧（未报告退化）。
  - 测试覆盖缺口：仅针对 token table 更新有单元测试，缺少端到端集成测试（模型加载 + 推理）。
- 影响：
  - 用户：可使用 LongCat-2.0-FP8 模型，在 8x B300 配置下 GSM8K 准确率 95.9%，吞吐约 1084 token/s。
  - 系统：配置文件解析逻辑增加对 LongCat 的自动识别，对非 LongCat 模型无影响（回退原逻辑）。
  - 团队：为后续添加类似 ngram 增强模型提供了可复用的 token 表更新和 indexer 扩展模式。
  - 风险标记：新模型引入，核心调度逻辑变更，DSA indexer 扩展，缺少端到端集成测试

关联脉络

- PR #30202 [Model] Support LongCat 2.0: 本 PR 的基础实现，该 PR 的 serving 修复和补充建立在 #30202 之上。
- PR #30042 [Ngram] Fix ngram token-table path for decode: 参考了该 PR 的 ngram token-table 更新路径实现，替代了之前 decode-only 的快速路径。