PR #22118 完整报告

sgl-project/sglang

[Score API] Add SequenceClassification Model support

合并时间：2026-04-08 16:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22118

执行摘要

- 一句话：为评分 API 添加序列分类模型支持，扩展多项目评分功能。
- 推荐动作：建议工程师精读此 PR，重点关注 score_and_pool 函数的实现，了解如何动态处理分隔符以支持多项目评分，以及 review 中的性能优化技巧（如避免 GPU-CPU 同步）。设计决策值得学习，特别是分类模型与生成模型的分发机制，以及 MIS 的高效打包策略。

功能与动机

根据 PR body，动机是扩展 /v1/score 端点以支持 SequenceClassification 模型，为分类模型提供评分能力，并添加多项目评分支持，提升 API 灵活性。具体表述为：'Extends the /v1/score endpoint to support SequenceClassification models (e.g., Qwen3ForSequenceClassification, Qwen2ForSequenceClassification, LlamaForClassification) in addition to the existing CausalLM path. For classification models, the scoring API returns pooled class logits from the model's classification head — no label_token_ids required.'

实现拆解

实现方案分为几个关键模块：1. 在 python/sglang/srt/layers/pooler.py 中新增 score_and_pool() 函数，动态查找分隔符位置并应用分类头，支持 MIS 和单项目评分路径。2. 修改 python/sglang/srt/managers/tokenizer_manager_score_mixin.py，根据模型类型（CausalLM 或 SequenceClassification）分发请求，处理 MIS 逻辑。3. 更新 python/sglang/srt/models/ 下的三个分类模型文件（llama_classification.py、qwen2_classification.py、qwen3_classification.py），将 forward 方法改为调用 score_and_pool()。4. 更新入口点文件（engine_score_mixin.py 和 http_server.py）的文档字符串，明确支持两种模型类型。5. 添加单元测试（test_pooler_score_and_pool.py）和 E2E 测试（test_score_classification.py），覆盖单项目、MIS 及边界情况。

关键文件：

- python/sglang/srt/layers/pooler.py（模块 layers）：新增 score_and_pool 函数，是处理多项目评分和单项目评分的核心逻辑，动态查找分隔符位置并应用分类头。
- python/sglang/srt/managers/tokenizer_manager_score_mixin.py（模块 managers）：修改以支持分类模型与生成模型的分发，处理多项目评分逻辑，是关键调度点。

- python/sglang/srt/models/llama_classification.py（模块 models）：更新 forward 方法使用 score_and_pool，集成序列分类模型支持，影响模型执行路径。
- test/registered/core/test_score_classification.py（模块 test）：新增 E2E 测试，验证序列分类模型的评分功能正确性，包括单项目和 MIS 路径。

关键符号：score_and_pool

评论区精华

review 中，gemini-code-assist[bot] 指出两个核心问题：1. 在 pooler.py 中，使用 GPU 张量 forward_batch.extend_seq_lens 迭代并调用 .item() 会导致 GPU-CPU 同步性能瓶颈，建议改用 CPU 列表 forward_batch.extend_seq_lens_cpu，代码已采纳此优化。2. 在 tokenizer_manager_score_mixin.py 中，使用 dim=0 进行 softmax 可能因张量形状变化导致错误，建议使用 dim=-1 以确保鲁棒性，代码已采纳。这些讨论聚焦于性能优化和正确性改进，已解决。

- GPU-CPU 同步性能优化 (performance): 建议改用 CPU 列表 forward_batch.extend_seq_lens_cpu 以避免同步，代码已采纳此优化。
- softmax 维度正确性 (correctness): 代码已采纳建议，使用 dim=-1 进行 softmax，避免潜在的正确性问题。

风险与影响

- 风险：技术风险包括：1. 核心路径变更：新函数 score_and_pool 逻辑复杂，分隔符动态查找可能引入 bug，尤其在处理边界情况（如无分隔符或分隔符在序列开头）时。2. 性能瓶颈：虽然 review 中优化了 GPU-CPU 同步，但 MIS 路径仍需监控前向传递开销。3. 兼容性影响：扩展评分 API 支持分类模型，可能影响现有 CausalLM 路径的兼容性，特别是在混合部署场景。4. 测试覆盖：新增测试较全面，但需确保 MIS 与单项目评分的集成测试覆盖所有模型类型。
- 影响：影响范围：对用户，评分 API 现在支持序列分类任务，提供更灵活的评分能力，无需手动处理 label_token_ids，提升易用性。对系统，API 功能增强，但引入新代码路径增加维护复杂性，需确保与现有 CausalLM 模型的无缝共存。对团队，新增测试和模块需要维护，并可能影响后续评分相关开发。影响程度中等，涉及核心评分逻辑和多个模块。
- 风险标记：核心路径变更，性能瓶颈风险，兼容性影响

关联脉络

- PR #20960 [Feature] Add token embedding overrides for sparse embedding replacement: 同样涉及评分 API 的增强功能，扩展了 API 能力，与本 PR 在评分系统演进上相关。
- PR #22405 [CICD] [prefill-only] Consolidate prefill-only model E2E tests: 涉及测试目录重构，与本 PR 中测试文件放置问题（如 Issue 评论中提到）相关，显示团队在统一测试基础设施。