Soft Thinking 推理演示

1. 项目介绍

本项目提供了一个比较 Soft Thinking 与标准 Chain-of-Thought (CoT) 推理方法的演示工具。Soft Thinking 是一种改进的思维链方法，通过在推理过程中保持多个思考路径并进行概率混合，可能提高大语言模型的推理能力。

2. 功能特点

- 支持使用相同模型对比 Soft Thinking 和标准 CoT 的推理效果
- 详细对比两种方法的生成时间、token 数量和回答长度
- 提供丰富的参数配置，支持自定义推理过程

3. 使用方法

最简单的方式是通过提供的 bash 脚本运行演示：

```
bash run_soft_thinking_demo.sh
```

4. 实验结果

1️⃣ **长度对比**

2️⃣ **token消耗对比**

3️⃣ **时间对比**

```
=== 性能对比汇总报告 ===

生成时间对比:
- 标准CoT: 63.12秒
- Soft Thinking: 19.72秒

Token数量对比:
- 标准CoT: 4096个token
- Soft Thinking: 869个token

回答长度对比:
- 标准CoT: 5668个字符
- Soft Thinking: 1350个字符

与标准CoT的相对比较:
- Soft Thinking vs 标准CoT:
  时间差异: -43.40秒 (-68.76%更短)
  Token差异: -3227个 (-78.78%更少)
  长度差异: -4318个字符 (-76.18%更短)
```

5. 参数说明

主要参数：

参数	说明	默认值
`--model_name`	模型名称或路径	"/root/autodl-tmp/Qwen3-1.7B"
`--prompt`	输入提示，不提供则使用预设数学问题	预设数学问题
`--max_generated_tokens`	生成的最大 token 数量	2048
`--temperature`	采样温度	0.6
`--top_p`	Top-p 采样概率	0.95
`--top_k`	Top-k 采样数量	30
`--think_end_str`	思考结束标记	`</think>`
`--max_topk`	Soft Thinking 的最大候选数量	10
`--num_gpus`	使用 GPU 数量	1

Soft Thinking 特有参数：

参数	说明	默认值
`--after_thinking_temperature`	思考后的采样温度	0.6
`--after_thinking_top_p`	思考后的 Top-p 采样概率	0.95
`--after_thinking_top_k`	思考后的 Top-k 采样数量	30
`--after_thinking_min_p`	思考后的 Min-p 采样概率	0.0
`--early_stopping_entropy_threshold`	早停熵阈值	0.0
`--early_stopping_length_threshold`	早停长度阈值	200

输出说明

程序会依次运行标准 CoT 和 Soft Thinking 两种方法，并显示：

1. 每种方法生成的完整回答
2. 生成信息（token 数量、结束原因、生成时间）
3. 两种方法的对比结果（时间、token 数量、回答长度等）

6. 关键参数详解

max_topk

`max_topk` 是 Soft Thinking 模式的核心参数，决定了模型在思考阶段会保留多少个可能的思考路径：

- **作用**：控制模型在思考过程中保持的候选路径数量
- **影响**：
  - 值越大，模型考虑的可能性越多，推理结果可能更全面
  - 值越小，计算效率越高，资源消耗越少
- **建议**：
  - 简单任务可设置为较小值（1-3）
  - 复杂推理任务可设置为较大值（5-10）

early_stopping_entropy_threshold

`early_stopping_entropy_threshold` 控制基于熵的早停机制：

- **原理**：当模型预测分布的熵低于此阈值时，触发提前停止生成
- **作用**：
  - 熵低表示模型对下一个token的预测非常确定
  - 可避免模型在已有明确答案时继续无效生成
- **取值范围**：
  - 0.0：禁用熵检查提前停止
  - 0.1-0.3：轻度提前停止
  - 0.3-0.5：中度提前停止
  - > 0.5：激进提前停止

early_stopping_length_threshold

`early_stopping_length_threshold` 控制基于长度的提前停止检查：

- **机制**：当生成的tokens数量达到此阈值时，开始检查是否可以提前停止生成
- **作用**：
  - 小值：更快开始检查停止条件，可能导致过早停止
  - 大值：允许模型生成更长内容后才检查停止条件
- **选择指南**：
  - 简单问题：5-20适合快速回答
  - 复杂问题：100-200允许更充分的思考

7. 优化建议

不同场景下的参数组合建议：

- **高效率模式**：
  - max_topk = 1
  - early_stopping_entropy_threshold = 0.3
  - early_stopping_length_threshold = 5
- **平衡模式**：
  - max_topk = 3-5
  - early_stopping_entropy_threshold = 0.2
  - early_stopping_length_threshold = 50
- **高质量模式**：
  - max_topk = 10
  - early_stopping_entropy_threshold = 0.1
  - early_stopping_length_threshold = 200

根据实际任务复杂度和可用计算资源调整这些参数，可以在推理质量和效率之间取得最佳平衡。