多模态PRM模型

浦博威

2025.04.14

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

VisualPRM: An Effective Process Reward Model for Multimodal Reasoning

Weiyun Wang[1,2], Zhangwei Gao[3,2], Lianjie Chen[4,2], Zhe Chen[5,2], Jinguo Zhu[2], Xiangyu Zhao[3,2], Yangzhou Liu[5,2], Yue Cao[5,2], Shenglong Ye[2], Xizhou Zhu[4,2], Lewei Lu[7], Haodong Duan[2], Yu Qiao[2], Jifeng Dai[4,2], Wenhai Wang[6,2] ✉

[1]Fudan University, [2]Shanghai AI Laboratory,
[3]Shanghai Jiaotong University, [4]Tsinghua University,
[5]Nanjing University, [6]The Chinese University of Hong Kong, [7]SenseTime Research

LLM-Math -> LLM-General -> MLLM-Math -?> MLLM-General

- 研究背景
- 相关工作
- 主要方法
- 实验结果

动机

challenges of adapting TTS for MLLMs：
(1) Lack of effective critic models.
(2) Lack of evaluation benchmarks for multimodal critic models.

directly evaluating critics under BoN settings poses two key issues：
1、使用BoN进行测试的时候，主要的消耗在于策略模型而非批判模型
2、批判模型收到多种策略模型的影响

- 研究背景
- 相关工作
- 主要方法
- 实验结果

相关工作

OpenAI 引出的PRM概念

PRM：过程奖励模型，是在生成过程中，分步骤，对每一步进行打分，是更细粒度的奖励模型。
ORM：结果奖励模型，是不管推理有多少步，对完整的生成结果进行一次打分，是一个反馈更稀疏的奖励模型。
参考Let’s Verify Step by Step中提供的思路看下PRM的作用：

首先，为了让模型有按步输出的能力，我们先通过一个按步骤回答的指令集，训练一个generator模型，模型不保证步骤一定是正确的，但能遵循指令按step1, step2,...格式输出。
然后，可以对上面的模型做N次采样（如best-of-N , Beam Search，lookahead Search方法等），并通过PRM对每个采样的每步推理做打分
最终，通过对每个步骤的打分拟合一个整体过程的打分，并按该整体打分选取打分最高的结果作为最终的答案。

Let’s Verify Step by Step (https://arxiv.org/pdf/2305.20050)

相关工作

PRM在DeepSeekR1里的失败
https://arxiv.org/pdf/2501.12948 4.2章节

First, it is challenging to explicitly define a fine-grain step in general reasoning.

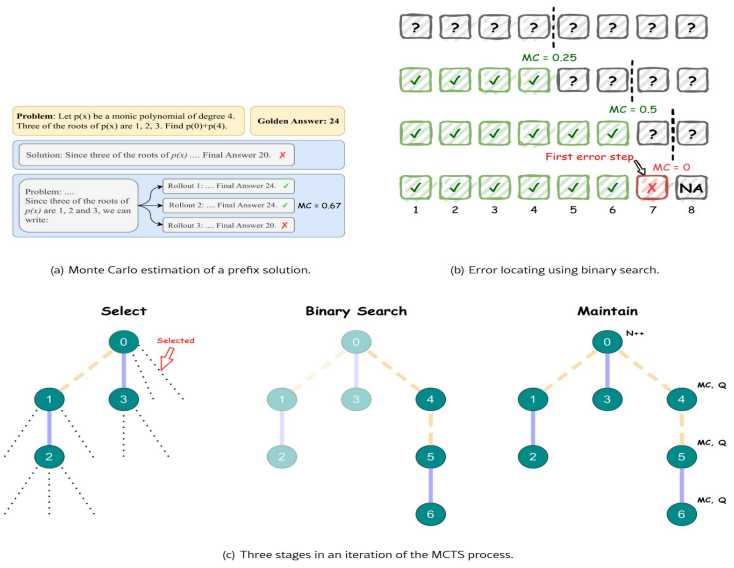
Second,determining whether the current intermediate step is correct is a challenging task. Automated annotation using models may not yield satisfactory results, while manual annotation is not conducive to scaling up.

Third, once a model-based PRM is introduced, it inevitably leads to reward hacking (Gao et al., 2022), and retraining the reward model needs additional training resources and it complicates the whole training pipeline.

相关工作

Improve Mathematical Reasoning in Language Models by Automated Process Supervision

蒙特卡洛（mc）收集高质量过程监督数据



- 研究背景
- 相关工作
- 主要方法
- 实验结果

解决方法：
1、构建数据集用于训练批判模型
2、构建benchmark

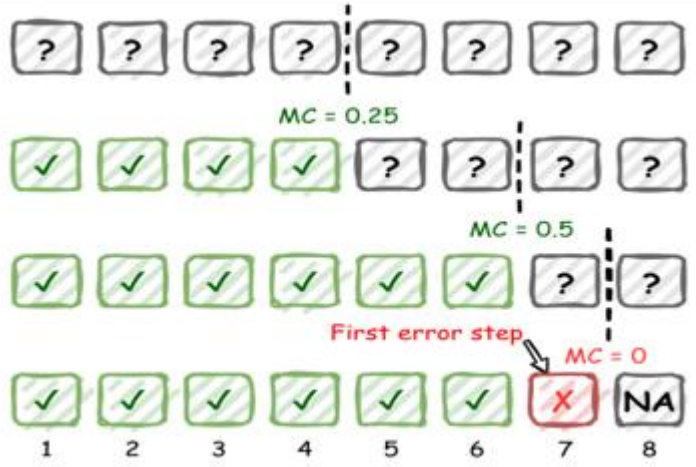
构建数据集

定义：
每条训练数据包含
1、图像
2、问题
3、每步的解答
4、每步的mc分数

mc分数是什么：

在模型输出多个结果中，正确的结果数量

为每对图像 - 问题样本选取 4 个解决方案，并将每个解决方案最多拆分为 12 个步骤。对于每个步骤，我们选取 16 个后续步骤，并根据这些后续步骤计算 m_i。最终得到的数据集包含约 40 万个样本和约 200 万个带有过程监督的步骤。



Process Supervision Generation. Given an image I, a question q, and a solution $s = \{s_0, s_1, \cdots, s_n\}$, we annotate the correctness of each step s_i using an automatic data pipeline. The key idea is to estimate the expected accuracy of given steps $s_{\leq i}$ based on Monte Carlo sampling. Specifically, the model is required to complete the solution as follows:

$$\tilde{s}_{>i} \sim M(\tilde{s}_{>i} \mid I, q, s_{\leq i}), \quad (1)$$

where $\tilde{s}_{>i}$ is the completion of $s_{\leq i}$. Besides, the expected accuracy of s_i is defined as:

$$mc_i = \frac{\text{num(correct completions)}}{\text{num(sampled completions)}}. \quad (2)$$

构建数据集

Before mc

Let's Verify Step by Step一文中，OpenAI详细描述了他们收集样本的方法

阶段1：冷启动，通过generator采集初版标注样本，对每个步骤采集多种表述给标注人员做标注，约5%的样本规模

阶段2：通过主动学习(active learning)标注难样本，提升模型对边界样本的学习能力。

而本文直接使用模型生成，猜测预训练和指令微调接环混合了这种数据

PRM训练

训练目标：

过程监督问题构建为一个多轮对话任务，有效利用多模态大语言模型（MLLMs）的生成能力。在第一轮对话中，会包含图像 I、问题 q 以及该问题解决方案的第一步 s0，随后每一轮都会给出一个新的步骤。要求模型在每一轮中对给定步骤的质量进行如下预测：

$$y_i \sim M(y_i \mid I, q, s_{\leq i}),$$

PRM训练

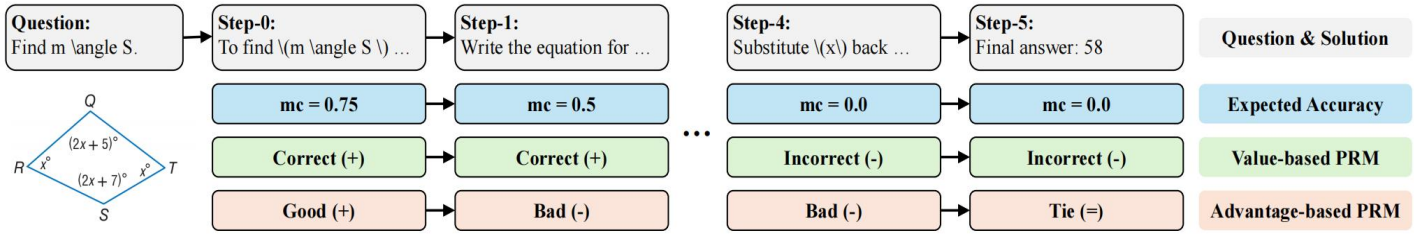


Figure 3. **Different modeling methods for PRMs.** PRMs are developed to estimate the quality of each step in a given solution. For value-based PRMs, the quality of a certain step is determined by its expected accuracy mc_i, where a step is considered correct if $mc_i > 0$. For advantage-based PRMs, the quality of a certain step is determined by the improvement of mc_i over mc_{i-1}, where a step is considered good if $mc_i - mc_{i-1} > 0$. During the training stage, the output space of PRMs is discretized into specific tokens, while during the inference stage, we compute the step score as the weighted sum of the generation probability for these discretized tokens.

2种类型的PRM

1.value-base:要求模型预测给定步骤的正确性，而非mc的精确分数。

2.adavantage-base: 某一步骤的质量由当前多准则指标mc相较于前一个指标的提升情况决定，这与强化学习中优势函数的定义类似。

推理中权重设置：

1.value-base: 仅仅{1,0}

2.adavantage-base {1,0,-1}(+,=,-)

VisualProcessBench

目标：基准测试要求模型找出给定解答中的所有错误步骤，而非仅仅找出第一个错误步骤。

标注方法：聘请了一组至少拥有大学学历的专业人员，由他们人工标注解答中每一步的正确性。具体而言，13人工作了3天，总计工作量为39人日。（可怕）

Statistics	Number
Total Samples	2866
- MMMU [90]	267
- MathVision [78]	712
- MathVerse [93]	1026
- DynaMath [99]	570
- WeMath [60]	291
Source Solutions	2866
- GPT-4o [58]	870
- Claude-3.5-Sonnet [4]	865
- QvQ-72B-Preview [72]	825
- InternVL2.5-78B [15]	306
Total Steps	26950
- Correct Steps	16585
- Incorrect Steps	7691
- Neural Steps	2674
Query Word Length Quartile	(22, 24, 50)
Response Word Length Quartile	(137, 193, 552)
Step Word Length Quartile	(13, 31, 67)
Number of Steps per Solution	9.4

Table 1. **Statistics of VisualProcessBench.**

- 研究背景
- 相关工作
- 主要方法
- 实验结果

实验

Model	MMMU	MathVista	MathVision	MathVerse-VO	DynaMath	WeMath	LogicVista	Overall
Proprietary Models								
GPT-4o [58]	70.7	60.0	31.2	40.6	34.5	45.8	52.8	47.9
Gemini-2.0-Flash [61]	69.9	70.4	43.6	47.8	42.1	47.4	52.3	53.4
Claude-3.5-Sonnet [4]	66.4	65.3	35.6	46.3	35.7	44.0	60.4	50.5
Open-source Models								
MiniCPM-V2.6-8B [89]	49.8	60.8	23.4	18.9	9.8	16.4	27.5	29.5
+VisualPRM	56.8	65.7	24.7	35.8	11.2	31.0	37.4	37.5
	+7.0	**+4.9**	**+1.3**	**+16.9**	**+1.4**	**+14.6**	**+9.8**	**+8.0**
Qwen2.5-VL-7B [7]	55.0	67.8	25.4	41.1	21.0	35.2	44.1	41.4
+VisualPRM	58.6	70.3	31.3	44.3	23.0	39.8	48.3	45.1
	+3.6	**+2.5**	**+5.9**	**+3.2**	**+2.0**	**+4.6**	**+4.2**	**+3.7**
InternVL2.5-8B [15]	56.2	64.5	17.0	22.8	9.4	23.5	36.0	32.8
+VisualPRM	60.2	68.5	25.7	35.8	18.0	36.5	43.8	41.2
	+4.0	**+4.0**	**+8.7**	**+13.0**	**+8.6**	**+13.0**	**+7.8**	**+8.4**
InternVL2.5-26B [15]	60.7	68.2	23.4	24.0	11.4	30.9	39.6	36.9
+VisualPRM	63.9	73.1	29.6	39.1	23.2	40.8	51.0	45.8
	+3.2	**+4.9**	**+6.2**	**+15.1**	**+11.8**	**+9.9**	**+11.4**	**+8.9**
InternVL2.5-38B [15]	63.9	71.9	32.2	36.9	20.0	38.3	47.9	44.4
+VisualPRM	69.0	73.9	35.2	46.7	30.5	46.2	53.7	50.7
	+5.1	**+2.0**	**+3.0**	**+9.8**	**+10.5**	**+7.9**	**+5.8**	**+6.3**
InternVL2.5-78B [15]	70.0	72.3	32.2	39.2	19.2	39.8	49.0	46.0
+VisualPRM	70.7	75.1	35.9	47.1	31.3	49.1	53.9	51.9
	+0.7	**+2.8**	**+3.7**	**+7.9**	**+12.1**	**+9.3**	**+4.9**	**+5.9**

Table 2. **Results on seven multimodal reasoning benchmarks.** MMMU [90] is a multidisciplinary reasoning benchmark. MathVista [50], MathVision [78], MathVerse [93], DynaMath [99], and WeMath [60] are mathematics benchmarks. For MathVerse, we report the performance on Vision-Only (VO) split. LogicVista [87] is a logical reasoning benchmark. Part of the results are collected from the OpenCompass leaderboard [19]. The overall score is the average score of the above benchmarks. By using VisualPRM as the critic model, existing open-source MLLMs achieve significant improvements in reasoning ability under the Best-of-8 evaluation strategy.

Future

应用：
用于RL训练，用于过程监督

局限性：
只能做数学题，并不是我们做的感知推理

我们任务迁移：
1、定义推理任务
2、定义步骤
解决这两个问题以后可以训练一个prm，用于RL训练