Small-Math-R1: 수학 추론을 위한 경량 RL 모델

Small-Math-R1: A Lightweight Reinforcement Learning Model for Mathematical Reasoning

이재건 [1],최장훈 [2]
[1] 경북대학교 데이터사이언스 석사과정
[2] 경북대학교 데이터사이언스 교수
leejken530@knu.ac.kr, jhchoi09@knu.ac.kr

요 약

본 연구는 경량 멀티모달 언어 모델 Qwen-2.5-VL 3B 를 기반으로, 수학 문제 해결에 특화된 소형 추론 모델 Small-Math-R1 을 제안한다. 모델 학습은 두 단계로 구성된다. 첫 단계에서는 수학 도메인 특화 멀티모달 데이터셋을 이용해 지도 미세조정(SFT) 을 수행하여 cold-start 문제를 완화한다. 두 번째 단계에서 Adjusted GRPO (Adjusted Group Reward Policy Optimization) 강화학습을 적용한다. Adjusted GRPO 는 기존 GRPO 의 objective 를 수정하여 (i) 분모를 생성된 총 토큰 수$(\sum_{i=1}^{G}|o_i|)$로 변경해 긴 추론을 장려하고, (ii) 참조 정책과의 KL 발산 페널티$-\beta D_{\mathrm{KL}}$를 제거해 정책이 보다 자유롭게 탐색하도록 유도한다. 또한 형식,길이,정답의 세 요소로 구성된 구조화된 보상 함수를 설계하여 모델의 행동을 정교하게 제어한다. 실험 결과,제안된 방법은 제한된 연산 자원 환경에서도 basemodel 인 Qwen-VL-3B 보다 훨씬 높은 정확도를 보여주었다.

1. 서론

최근 멀티모달 대형 언어 모델(MLLM)은 텍스트, 이미지 등 다양한 정보를 통합 처리하는 능력이 발전했지만[1], 복잡한 기호 조작, 다단계 추론, 시각 요소(도형, 그래프) 해석이 필요한 수학 문제 해결에는 여전히 어려움을 겪는다. 특히, 엄격한 출력 형식(예: 사고 과정과 답안 분리)을 준수하며 제한된 자원으로 운용 가능한 경량 수학 특화 MLLM 개발은 미흡하다. 수학 문제 해결은 논리적 단계 제시가 중요하며 특정 출력 형식을 요구한다. 기존 강화학습 기반 미세조정(RLHF) 기법(PPO[3], DPO[4], GRPO[5] 등)은 일반적인 응답 품질 개선에는 효과적이나, 표준 GRPO[6]의 균등 가중치 부여 및 KL 발산 제어 방식은 수학 문제의 긴 추론 연쇄 생성과 창의적 문제 해결 전략 탐색을 제약할 수 있다.

본 연구는 이러한 한계를 극복하기 위해 Qwen-2.5-VL 3B[7] 기반의 경량 수학 추론 모델 Small-Math-R1 을 제안한다. 주요 기여는 다음과 같다: (1) 토큰 길이에 비례한 가중치 부여 및 KL 페널티 제거로 긴 추론과 탐색을 장려하는 **Adjusted GRPO** 제안. (2) 수학 문제 해결 특성(형식, 길이, 정답률)을 반영한 **구조화된 보상 함수** 설계. (3) 3B 파라미터 소형 모델로도 멀티모달 수학 문제 해결에서 높은 성능과 형식 일관성을 달성함을 입증 (기존 Qwen-2.5-VL 3B 대비).

2. 관련연구

GPT-4V, Gemini, LLaVA[2], Qwen-VL[6] 등 기존 MLLM 은 뛰어난 일반 시각-언어 이해 능력을 보이나, 대부분 규모가 크고 수학 추론에 최적화되지 않았다. PPO[3], DPO[4], GRPO[5] 등 RLHF 기법은 모델 정렬에 기여했지만, 수학 문제의 긴 추론이나 탐색에는 제약이 따를 수 있다. Minerva, MathCoder[7] 등 수학 특화 LLM 도 등장했으나, 주로 텍스트 기반이거나 대규모 모델이며, 멀티모달 입력을 처리하고 엄격한 출력 형식을 따르는 경량 모델 연구는 부족하다. Small-Math-R1 은 이러한 경량 멀티모달 수학 추론 및 형식 준수라는 틈새를 목표로 한다.

3. 방법론

3.1 모델 아키텍처

본 연구는 30 억 파라미터 경량 MLLM 인 Qwen-2.5-VL 3B[7]를 기반으로 합니다. 텍스트와 시각 정보(이미지 프레임)를 함께 처리하여 수학적 추론 과정이 포함된 텍스트 답변을 생성하는 이 모델은, 효율성과 성능의 균형 덕분에 수학 추론에 적합하다.

3.2 학습 파이프라인 (Training Pipeline)

Small-Math-R1 의 학습은 지도 미세조정(SFT)과 Adjusted GRPO 기반 강화학습(RL) 두 단계로 구성된다

3.2.1 (1 단계): 지도 미세조정(SFT)

SFT 단계의 목적은 사전 학습된 Qwen-2.5-VL 3B 모델을 수학 도메인에 적응시키고, 목표 출력 형식(예:<think>...</think><answer>...</answer>)에 익숙해지도록 하는 것이다. 이를 위해 수학 문제(텍스트)와 관련 멀티모달 입력(이미지), 그리고 상세한 풀이 과정(<think> 태그 내부) 및 최종 정답(<answer> 태그 내부)으로 구성된 고품질 데이터셋을 사용한다. 이 단계는 모델이 수학 용어, 기호, 그리고 기본적인 문제 해결 패턴을 학습하도록 돕고, 후속 RL 단계에서의 cold-start 문제를 완화하여 학습 효율성을 높인다. 학습은 표준 언어 모델링 손실(cross-entropy loss)을 최소화하는 방향으로 진행된다.

3.2.2 (2 단계): Reinforcement Learning with Adjusted GRPO
표준 GRPO 는 샘플 $o_{i=1}^{G}$ 에 대해 다음과 같은 PPO-Clip objective 를 사용한다.

$$J_{GRPO}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)(\hat{A}_{i,t}),\text{clip}(r_{i,t}(\theta),1-\varepsilon,1+\varepsilon)(\hat{A}_{i,t})\right) - \beta D_{KL}(\pi_\theta \parallel \pi_{\text{ref}})$$

여기서 $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|o_{i,<t})}$ 은 확률비율이고, $(\hat{A}_{i,t}) = \frac{R_i - \text{"mean"}\left(\{R_j\}_{j=1}^{G}\right)}{\text{"std}}$, $\left(\{R_j\}_{j=1}^{G}\right)$ 는 보상정규화이다.

Adjusted GRPO Objective
수학 문제 해결에서는 상세하고 긴 추론 과정이 중요하며, 때로는 기존 풀이 방식과 다른 창의적인 접근이 필요하다. 이를 반영하기 위해 Adjusted GRPO 는 다음과 같이 수정된다:

$$\mathcal{J}_{AJ}(\theta) = E(q,a)\sim\mathcal{D},\{o_i\} \sim \pi_{\text{old}}\left[\frac{1}{\sum_{j=1}^{G}|o_j|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\,(\hat{A}_{i,t}),\text{clip}(r_{i,t}(\theta),1-\varepsilon,1+\varepsilon)(\hat{A}_{i,t})\right)\right]$$

(1) **토큰 길이 기반 가중치 (Token Length Based Weighting):** 샘플별 균등 가중치 대신, 생성된 총 토큰 수에 기반한 가중치를 적용한다. 이는 긴 응답(상세한 추론)을 생성한 샘플 그룹에 더 큰 학습 비중을 두도록 한다. objective 함수의 분모를 G 대신 $\sum_{j=1}^{G}$ $|o_j|$로 변경한다.
(2) **KL 페널티 제거 (Removal of KL Penalty):** 참조 정책 π_{old}로부터의 이탈을 제어하는 $-\beta D_{\text{KL}}$ 항을 제거한다. 이는 정책 π_θ가 SFT 모델의 분포에서 더 자유롭게 벗어나 새로운 추론 경로와 해결 전략을 탐색할 수 있도록 허용한다. 잠재적인 학습 불안정성은 정교하게 설계된 보상 함수와 PPO 스타일의 클리핑 메커니즘을 통해 완화한다.

3.3 구조화된 보상함수(Structured Reward Function)
세 개의 부분 보상:
- **형식 보상** $r_0 \in [0,1]$:태그 형식 준수 시 1 틀리면 0
- **길이 보상** $r_1 \in [0,1]$:목표 길이 L_{target}에 따라 $r_1 = \min\left(1,\frac{\text{len}(t)}{L}\right)$
- **정답 보상** $r_2 \in 0,1$ – 정답 일치 시 1.

형식/길이 통합 점수 FR와 최종 보상 R는 다음과 같이 정의한다.

$$FR = r_0 + r_{\{ratio\ \}} * r_1, AR = r_2$$

- **Total Reward (R)**

$$R = \begin{cases} AR + FR, & \text{if } FR > 0 \text{ and } AR = r_2 \\ -FR, & \text{if } FR > 0 \text{ and } AR = 0 \\ -(r_0 + r_1 + r_2), & \text{if } FR = 0 \end{cases}$$

이 설계는 (i) 먼저 형식을 맞추고 (ii) 정답을 찾으며 (iii) 가능한 한 자세한 추론을 씀으로써 최고 보상을 얻도록 유도한다.

4. 실험 (Experiments)

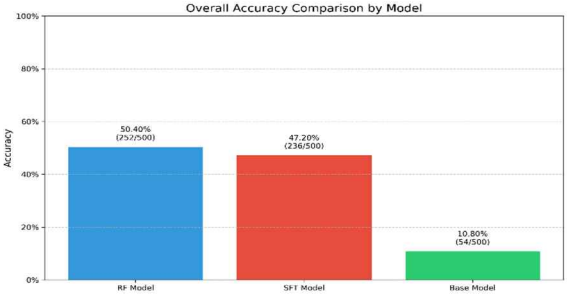
본 장에서는 Small-Math-R1 의 성능을 **Math-500** 데이터셋에서 평가할 예정이다. Math-500 은 난이도별 500 문제로 구성된 대표적 수학 벤치마크로, 텍스트 설명과 그림/표 등 시각 자료가 포함된다

4.1 데이터셋 (Dataset)
- **Math-500**: 산술 · 대수 · 기하 · 확률/통계 등 네 영역으로 구성된 500 문제.
- 각 문제는 문제문, 관련 이미지(선택)와 정답 및 해설(형식)으로 이루어져 있다.

4.2 비교 모델 (Baselines)

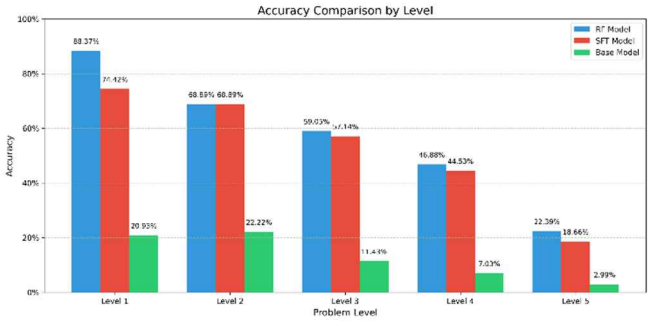
모델	파라미터	비고
Small-Math-R1 (Ours)	3B	Adjusted GRPO + 구조화 보상
Qwen-2.5-VL SFT	3B	Base + SFT
Qwen-2.5-VL 3B	3B	Base model

4.3 결과 및 분석
4.3.1 전체 정확도 비교



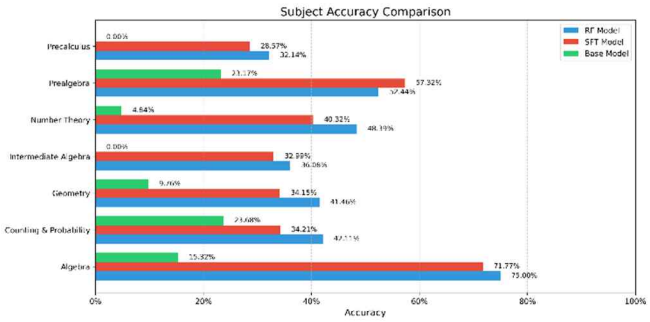
Math-500 평가에서 제안 모델(Small-Math-R1)은 50.4% (252/500 문제)의 정확도를 달성, 기본 모델(Qwen-2.5-VL 3B, 10.8%, 54/500 문제) 대비 큰 성능 향상을 보였습니다. 이는 구조화된 보상 기반 강화학습의 효과를 입증한다.

4.3.2 난이도별 정확도 비교



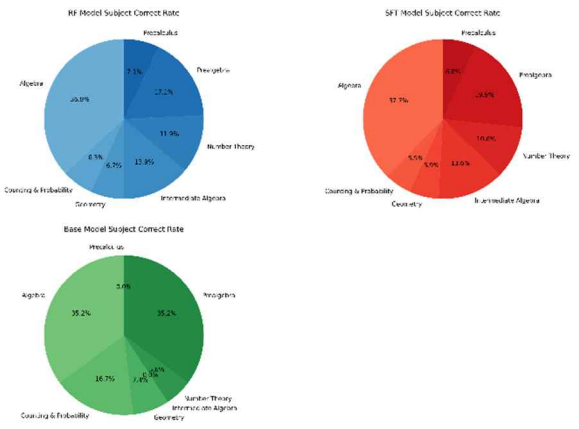
RF 모델은 전 난이도에서 Base 모델 대비 우수했으며 (Level 1: 88.37% vs 20.93%, Level 5: 22.39% vs 2.99%), 난이도 증가에 따른 정확도 감소 폭도 완만하여 안정적인 추론 능력을 입증했다.

4.3.3 주제별 정확도 비교



RF 모델은 Algebra(75.0% vs 15.3%), Number Theory(48.4% vs 4.8%) 등 대부분 수학 영역에서 Base 모델 대비 현저히 높은 정확도를 기록하였고, Base 모델이 풀지 못한 영역에서도 RF 모델은 준수한 성능을 보여 다양한 수학 주제에 걸쳐 편향 없는 효과적인 학습을 입증했다..

4.3.4 분포 기반 정성 비교



파이 차트는 RF 모델의 정답 분포가 다양한 주제와 난이도에 걸쳐 비교적 고르게 분포되어 있음을 보여준다(Algebra 에 대한 높은 집중도를 보이긴 하나, 다른 주제에서도 일정한 성능을 유지). 반면, Base 모델은 정답이 특정 주제(Algebra, Precalculus 등)와 낮은 난이도(Level 2)에 편중되어 있으며, 다수의 주제에서 zero-correct 현상(정답률 0%)이 나타나는 경향을 보인다.

5. 결론 (Conclusion)

본 연구는 경량 모델 Small-Math-R1, Adjusted GRPO, 구조화 보상 함수를 제안하여 엄격한 형식과 높은 정답률을 갖춘 수학 추론 모델을 개발하였으며, 향후 추가 RL, 부분 보상, 도메인 확장, RL 알고리즘 고도화를 통해 성능과 안정성을 더욱 향상시킬 것이다.

6. 감사의 말(Acknowledgement)

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2021R1C1C2095450, RS-2023-00242528)과 정보통신기획평가원 -대학 ICT 연구센터사업 (ITRC)의 지원 (IITP-2024-RS-2024-00437756)을 받아 수행된 연구임.

참고문헌

[1] Zhang, Boqiang; Li, Kehan; Cheng, Zesen; Hu, Zhiqiang; Yuan, Yuqian; **et al. "VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding."** *arXiv preprint* arXiv:2501.13106, 2025. DOI: 10.48550/arXiv.2501.13106

[2] Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. "**Visual Instruction Tuning**." arXiv preprint arXiv:2304.08485.

[3] Schulman, John; Wolski, Filip; Dhariwal, Prafulla; Radford, Alec; Klimov, Oleg. **"Proximal Policy Optimization Algorithms."** arXiv preprint arXiv:1707.06347, 2017.

[4] Rafailov, Rafael; Sharma, Archit; Mitchell, Eric; Ermon, Stefano; Manning, Christopher D.; Finn, Chelsea**. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model."** arXiv preprint arXiv:2305.18290, 2023.