순위 가중 보상을 활용한 대규모 언어 모델의 강화학습 기반 최적화

장준영°, 김기응
한국과학기술원
{junjang99, kekim}@kaist.ac.kr

Rank-Weighted Reward Optimization for Reasoning in Language Language Models via Reinforcement Learning

요 약

수학 문제 풀이와 같은 추론 중심 과제에서의 언어 모델 최적화를 위해 기존 GRPO 방식의 한계를 개선한 critic-free 정책 학습 알고리즘을 제안한다. 기존 방식은 그룹 내 top-1 응답에만 학습 신호가 집중되고, 나머지 샘플은 그래디언트에 기여하지 못해 학습 분산이 증가하고 샘플 효율성이 저하되는 문제가 있다. 이를 해결하기 위해 본 논문에서는 보상 기반 응답 순위를 부드럽게 확률화하여 모든 샘플이 기여할 수 있도록 하는 학습 구조를 설계하였다.

1. 서론

대규모 언어 모델(LLM)은 다양한 자연어 처리 과제를 해결하는 데 있어 강력한 성능을 보이며, 특히 수학 문제 풀이, 논리적 질의응답, 계획 수립 등 추론 능력을 요구하는 복잡한 문제에서도 핵심 도구로 자리잡고 있다 [1, 2]. 이러한 추론 능력을 정밀하게 개선하기 위해서는 단순한 언어 생성 과정 이상으로 긴 추론 과정(CoT)을 학습하는 방식이 요구된다 [3]. 최근에는 지도학습 기반 미세조정(SFT)에 이어 강화학습 기반의 사후 학습이 주목을 받고 있다. [4]

기존 강화학습 기반 LLM 최적화는 대부분 가치 함수를 사용하는 actor-critic 구조 [5]를 사용한다. 그러나 이러한 방법은 가치 함수 예측의 불안정성을 수반한다. 이를 해결하기 위해 최근 제안된 Groupwise Relative Policy Optimization(GRPO)은 가치함수 없이도 학습이 가능한 새로운 방향을 제시한다 [6]. GRPO는 같은 입력에 대해 생성된 여러 개의 후보 응답을 그룹으로 묶고, 각 응답에 대해 부여된 보상을 z-score 정규화를 통해 상대적 어드밴티지로 변환한 후 정책을 업데이트한다.

하지만 대부분의 샘플이 음수 어드밴티지를 받아 그래디언트가 억제되고 실질적으로 가장 높은 보상을 받은 샘플에만 양수 그래디언트가 집중되어 학습 분산이 커지며 그룹 내 순위 정보가 존재함에도 불구하고 그 구조를 온전히 활용하지 못한다는 단점이 있다.

이에 본 논문에서는 GRPO의 이러한 한계를 보완하기 위해 대규모 언어 모델 사후 학습 방식을 제안한다. 제안 기법은 (1) 보상 기반 순위를 부드럽게 확률화하여 모든 샘플이 기여할 수 있도록 하고, (2) 수학 문제와 같은 추론 중심 태스크에 초점을 맞추어 성능 향상을 도모한다.[1]

1) 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. RS-2024-00343989, 사회적, 윤리적 학습을 위한 데이터 특성 및 생성 AI 모델의 윤리성 향상 연구, No.2020-0-00940, 안전한 강화학습 원천 기술 개발 및 자연어 처리에의 응용)

2. 연구 배경

본 논문에서 해결하고자 하는 문제는 대규모 언어 모델은 수학 문제 풀이이다. 이는 주어진 문제에 대해 추론 과정을 거쳐 정답을 도출하는 과정이다. 대규모 모델에게 추론 능력을 갖게 하기 위해 강화학습을 활용하는데 GRPO는 가장 널리 쓰이는 알고리즘 중 하나이다. 본 논문에서는 GRPO의 문제점을 제시하고 개선하여 학습 분산을 줄이고 수렴 속도를 빠르게 하는 알고리즘을 제안한다.

GRPO GRPO는 가장 널리 쓰이는 critic-free 강화학습 방식으로, 가치 네트워크 없이 여러 개의 후보 응답에 대해 그룹 내 상대적 보상 차이만을 활용하여 정책을 최적화한다. 입력 prompt 하나에 대해 생성된 응답 집합 $\{y_1, y_2, \ldots, y_K\}$에 대해, 각 응답 y_i에 대응하는 보상 $r_i \in \mathbb{R}$를 부여하고, 이를 다음과 같이 그룹 내 정규화된 어드밴티지로 변환한다:

$$\hat{A}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$$

이 방식은 가치 함수 없이도 보상 신호만으로 정책을 업데이트할 수 있는 장점이 있다.

3. 본론

기존 GRPO는 각 샘플의 보상 에 대해 그룹 평균 및 표준편차를 이용한 어드밴티지를 계산하고, 이를 곧바로 정책 업데이트에 활용한다. 이는 평균이 0으로 정규화되기 때문에 대부분의 값이 음수가 되고, 가장 보상이 큰 하나의 응답만이 양수 그래디언트를 받게 될 확률이 높다.

$$\hat{A}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})} \quad w_i = \frac{\exp(\hat{A}_i/\tau)}{\sum_j \exp(\hat{A}_j/\tau)}$$

여기서 τ는 하이퍼파라미터로 낮을수록 상위 샘플에 집중하고 높을수록 균등 분포에 가까워진다. 이때 최종 어드밴티지는 단일 샘플의 $\hat{A}_i$가 아닌 다음과 같은 가중 보상으로 정의된다:

$$\tilde{A}_i = w_i \cdot r_i$$

이를 통해 모든 샘플이 그 중요도에 비례하여 그래디언트에 기여하게 되고 샘플 효율성과 학습 안정성을 동시에 확보할 수 있다. 최종적인 손실 함수는 다음과 같다.

$$\begin{aligned}\mathcal{L}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\Bigg[\\ \min\Big(\rho_{i,t}\cdot\tilde{A}_i,\ \text{clip}(\rho_{i,t},1-\epsilon,1+\epsilon)\cdot\tilde{A}_i\Big) \\ -\beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg]\end{aligned} \tag{1}$$

수식 1에서 $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$로 중요도 샘플링 비율을 나타내고 두 번째 항은 KL 정규화으로 기존의 정책에서 크게 벗어나는 것을 제한한다.

Algorithm 1 SoftRank-Guided Policy Optimization

Require: Policy π_θ, old policy π_{old}, reward function R, temperature τ, clipping range ϵ

for each prompt x **do**
  Generate K responses $\{y_1, \ldots, y_K\} \sim \pi_{\text{old}}$
  Compute rewards $\{r_1, \ldots, r_K\}$
  Normalize: $s_i \leftarrow (r_i - \mu)/\sigma$
  SoftRank weights: $w_i \leftarrow \exp(s_i/\tau)/\sum_j \exp(s_j/\tau)$
  Soft advantage: $\tilde{A}_i \leftarrow w_i \cdot r_i$
  for each token (a_t, s_t) in chosen response **do**
    $\rho_t \leftarrow \pi_\theta(a_t|s_t)/\pi_{\text{old}}(a_t|s_t)$
    $\mathcal{L}_t \leftarrow -\min\Big(\rho_t \cdot \tilde{A}, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\cdot\tilde{A}\Big)$
  end for
end for
Aggregate $\mathcal{L}$ and apply KL regularization
Update θ via gradient descent

4. 실험

본 연구에서는 제안한 알고리즘의 효과를 수학 문제 풀이 과제에 적용하여 검증하였다. 알고리즘 1의 과정을 대표적인 수학 추론 벤치마크인 GSM8K [7]를 통해 실험하였으며, 모델로는 Qwen 2.5 3B Instruct [8]를 선택하였다. 미세조정은 QLoRA [9] 기반 4비트 양자화 방식으로 수행하였으며, 학습 효율성과 속도를 극대화하기 위해 Unsloth 라이브러리를 [10] 사용하고, Flash Attention 2 [11]와 vLLM을 [12] 통해 추론 속도를 개선하였다. 생성 토큰 수는 최대 512로 제한하였고, 어드밴티지 계산 과정에서 사용하는 하이퍼파라미터 τ는 0.5로 고정하였다.

모든 응답은 추론 부분과 정답 부분이 명시적으로 구분되도록 `<reasoning>` 및 `<answer>` 태그로 감싸서 학습에 사용하였다. 이 포맷은 DeepSeek에서 [2] 사용한 스타일을 참고한 것으로, 추론 과정과 최종 답변을 모델이 명확하게 구분하여 학습할 수 있도록 한다.

비교 실험은 기존의 GRPO 방식과 본 논문에서 제안한 방식 간의 성능 차이를 중심으로 진행되었다. 두 실험 조건 모두 동일한 초기 모델 가중치, 동일한 데이터셋 분할, batch size, 학습 step 수를 설정하였으며, 공정한 비교를 위해 동일한 보상 함수와 샘플 수를 유지하였다. 기존 GRPO는 각 응답에 대해 z-score 정규화를 수행하는 반면 본 논문의 알고리즘은 보상을 기반으로 소프트맥스를 통해 부드러운 순위 가중치를 계산한 후, 이 가중합을 어드밴티지로 사용한다.

평가는 정책의 평균 보상을 중심으로 안정적인 학습이 진행되는지를 중심으로 한다.

결과 그림 1은 학습 과정 동안의 평균 보상 추이를 비교한 결과이다. GRPO와 제안한 SoftRank-GRPO 모두 초기 단계에서는 유사한 증가 양상을 보였으나, 약 150 step 이후부터 두 방법 간 뚜렷한 차이가 나타났다. 우리의 알고리즘은 약 200 step을 기점으로 빠르게 보상이 상승하며, 기존 GRPO보다 약 100 step 빠르게 학습이 수렴하는 양상을 보였다. 이후 단계에서도 평균 보상이 더 높은 수준에서 유지되며, 진동이 적고 보다 안정적인 추세를 보인다. 이는 제안한 알고리즘이 그룹 내 모든 샘플의 정보를 부드럽게 활용함으로써 그래디언트 분산을 줄이고 더 빠른 정책 개선을 유도한 결과로 해석할 수 있다. 특히 기존 GRPO의 경우 top-1 응답에만 reward가 집중되어 학습 속도가 제한되는 반면, SoftRank-GRPO는 전체 응답 분포를 반영하므로 보상 구조에 대한 학습이 효율적으로 이뤄진다. 또한 어드밴티지 구조를 부드럽게 설계했기 때문에, 학습의 안정성이 동시에 확보되었음을 확인할 수 있다. 표 1에서 확인할 수 있다시피 최종 학습 이후에도 더 높은 성능을 보임을 확인할 수 있다. 이러한 결과는 제안한 알고리즘이 수학 추론 태스크와 같은 구조적 추론 문제에서 빠른 수렴과 안정적인 학습을 달성할 수 있음을 보여준다. 대규모 학습에서는 이러한 장점이 더 크게 나타나고 비용적 측면에서도 효과적일 것으로 기대된다.

Method	**GRPO**	**SR-GRPO (ours)**
Acc (%)	72.4	**75.6**

표 1: 각 방법에 대한 정확도(Acc, %) 비교. 제안한 방법이 가장 높은 성능을 기록함.

그림 1: 실험 결과. 학습 step에 따른 평균 보상 비교. 제안한 알고리즘은 기존 GRPO보다 더 빠르게 수렴하며, 전반적으로 더 높은 reward를 안정적으로 유지한다.

5. 결론

본 연구에서는 기존 GRPO 방식이 지니는 top 1 응답에 대한 그래디언트 집중 현상과 샘플 비효율성 문제를 해결하기 위한 새로운 critic-free 최적화 알고리즘을 제안하였다. 제안한 알고리즘은 보상 값의 순위 정보를 부드럽게 확률화하는 방식을 통해 그룹 내 모든 응답이 정책 업데이트에 기여할 수 있도록 구성되었으며, 이를 기존 구조와 결합하여 정책 변화에 대한 안정성 또한 함께 확보하였다. 이러한 방식은 강화학습 기반 파인튜닝 과정에서 그래디언트의 흐름을 보다 부드럽고 안정적으로 만들며, 특히 추론 중심 태스크에서 효과적인 학습 구조를 제공한다. GSM8K 수학 문제 풀이 데이터셋을 기반으로 진행한 실험에서 제안한 방법은 기존의 그룹 기반 강화학습 방식에 비해 더 빠른 수렴 속도와 높은 보상 안정성을 보였다. 초기 학습 단계부터 평균 reward가 빠르게 상승하고, 전체 학습 기간 동안 높은 수준에서 유지되며, 전반적인 학습 효율성과 안정성이 모두 향상됨을 확인하였다. 이는 제안한 구조가 응답 간의 상대적 품질 정보를 효과적으로 반영하고, 가치 함수 없이도 충분한 학습 신호를 제공할 수 있음을 의미한다. 향후 연구에서는 본 알고리즘을 다양한 추론 태스크 및 언어 생성 과제에 확장하고 대규모 학습 환경에서도 그 안정성과 일반화 가능성을 평가할 수 있을 것으로 기대한다.

참고 문헌

[1] OpenAI, "Learning to reason with llms," 2024. Available at https://openai.com/research/learning-to-reason-with-llms.

[2] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, brian ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.

[4] E. Zelikman, G. R. Harik, Y. Shao, V. Jayasiri, N. Haber, and N. Goodman, "Quiet-STar: Language models can teach themselves to think before speaking," in *First Conference on Language Modeling*, 2024.

[5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[6] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," 2024.

[7] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[8] Q. Team, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[9] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[10] D. Han, "Unsloth: Fast and memory-efficient fine-tuning for llms." https://github.com/unslothai/unsloth, 2023.

[11] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," in *The Twelfth International Conference on Learning Representations*, 2024.

[12] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "vllm: A high-throughput and memory-efficient inference and serving engine for llms." https://github.com/vllm-project/vllm, 2023.