Hindsight PPO : RLHF를 위한 토큰 수준 기여도 할당

황두환[0], 김기응
한국과학기술원
{howdydoo0711, kekim}@kaist.ac.kr

Hindsight PPO : Token-Level Credit Assignment for RLHF

Doo Hwan Hwang[0], Kee-Eung Kim
Korea Advanced Instituite of Science and Technology
{howdydoo0711, kekim}@kaist.ac.kr

요 약

인간 피드백 기반 강화학습(Reinforcement Learning from Human Feedback, RLHF)은 대규모 언어 모델(Large Language Models, LLMs)을 인간 선호에 정렬시키는 데 널리 사용된다. 그러나 기존 RLHF 방법은 주로 문장 수준 보상에 의존하기 때문에, 개별 토큰이 최종 출력에 미치는 기여를 정교하게 평가하지 못한다. 이로 인해 정책 업데이트 과정에서 신용 할당이 부정확해지고, 높은 분산과 낮은 학습 효율이 초래된다. 본 논문에서는 이러한 한계를 완화하기 위해, 사후적 분포(Hindsight distribution)로부터 유도된 어드밴티지 함수를 통해 토큰별 기여도를 추정하는 Hindsight PPO(HPPO)를 제안한다. HPPO는 출력 임베딩을 이용해 사후적 분포를 근사하고, 하위 문장과 전체 문장사이의 의미적 거리를 기반으로 한 에너지 기반 모델링을 통해 문맥 인식형 토큰 수준 피드백을 제공한다. 실험결과, HPPO는 인간 선호와의 정렬 성능을 향상시키는 동시에 일반화 능력을 유지하며, 기존 RLHF방법론들보다 우수한 성능을 보인다.

1. 서 론 (Introduction)

대규모 언어 모델은 수학적 추론, 의사결정, 그리고 자연어 생성 전반에서 뛰어난 성능을 보여주고 있다. 그러나 성능 향상과 함께, 모델이 인간에게 유용하고 무해하며 정직한 응답을 생성하도록 정렬하는 문제는 더욱 중요해지고 있다[1]. 이를 위해 강화학습 기반 인간 피드백(RLHF)[2]과 직접 선호 최적화(DPO)[3]와 같은 방법이 널리 활용되고 있다. 하지만 기존 방법들은 주로 생성 완료 이후 문장 수준에서 보상을 부여하기 때문에, 최종 보상을 개별 토큰에 정확히 할당하기 어렵다. 이는 강화학습의 희소하고 지연된 피드백 문제와 유사하며, 토큰 수준 생성 과정과 문장 수준 보상 사이의 불일치로 인해 비효율적인 신용 할당, 느린 수렴, 그리고 불안정한 정책 업데이트를 초래한다[4]. 그 결과 모델은 참조 정책에서 과도하게 벗어나거나, 반대로 지나치게 보수적인 출력으로 붕괴될 수 있다. 또한 기존 접근법은 신용할당을 위해 추가적인 모델(Value model)을 도입하여 계산 비용을 증가시키거나, 어텐션 가중치 또는 균일한 어드밴티지 전파에 의존함으로써 실제 토큰별 기여도 차이를 충분히 반영하지 못한다.

본 논문에서는 이러한 한계를 해결하기 위해 Hindsight PPO (HPPO)를 제안한다. HPPO는 사후적 분포에 기반해 토큰 수준 어드밴티지를 추정함으로써, 추가적인 언어 모델 없이도 각 토큰의 기여도를 보다 정밀하게 평가한다. 이를 통해 출력 임베딩 공간에서의 의미론적 관계를 활용한 문맥 인식형 신용 할당이 가능해지며, 보다 안정[1)]적이고 효율적인 정렬 학습을 가능하게 한다.*

2. 관련 연구 (Related Work)

2.1. RLHF에서의 토큰 수준 피드백

RLHF의 핵심 난제 중 하나는 보상이 주로 문장 수준에서 주어지기 때문에, 개별 토큰이 최종 결과에 어떻게 기여했는지를 정교하게 평가하기 어렵다는 점이다. 이러한 희소하고 지연된 피드백은 비효율적인 신용 할당, 느린 수렴, 그리고 차선의 정책 업데이트로 이어질 수 있다. 이를 해결하기 위해 일부 연구는 외부 LLM을 활용해 토큰 수준 보상을 생성하였다. 예를 들어, DRLC[5]는 각 토큰의 기여 방향을 판별해 이산 보상을 할당하고, TLCR[6]은 대규모 LLM과 편집 기반 분석을 통해 연속적인 토큰 선호를 추정한다. 그러나 이러한 방법들은 반복적인 LLM 추론을 필요로 하므로 계산 비용이 크다. 다른 계열의 방법들은 외부 LLM 없이 보상 모델이나 정책 최적화 구조를 활용한다. ABC[7]는 보상 모델의 어텐션 정보를 이용해 토큰별 기여도를 재분배하지만, 어텐션 가중치가 실제 인과적 기여를 충실히 반영하지 못할 수 있다. 한편 GRPO와 Dr.GRPO는 배치 내 상대적 보상을 이용해 문장 수준 어드밴티지를 추정함으로써 추가 가치 모델 없이 효율적인 학습을 가능하게 하지만, 동일한 어드밴티지를 모든 토큰에 일괄적으로 부여하여 문맥 의존적인 기여도 차이를 반영하지 못한다[8].

*) 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임
(No. RS-2022-II220311,일상적 물건들의 다접촉 로봇 조작을 위한 목적 지향 강화학습 기술개발)

3. 배경지식 (Preliminaries)

3.1. MDP를 이용한 언어생성

LLM 기반 RLHF는 M=(S,A,P,r,ρ,H)로 정의되는 MDP로 볼 수 있으며, 목표는 인간 선호에 부합하는 시퀀스를 생성하도록 정책을 최적화하는 것이다. 자기회귀 생성에서 상태 st는 초기 프롬프트 x와 이전까지 생성된 토큰들의 prefix,

$$s_t = (s_0, a_0, a_1, \ldots, a_{t-1})$$

로 주어지며, 다음 토큰은 at∼πθ(·|st)에 따라 샘플링된다. 생성은 종료 상태 sH에 도달하면 끝나고, 보상 모델은 완성된 시퀀스에 대해 스칼라 보상을 부여한다. 정책은 PPO와 같은 방법으로 최적화되며, 참조 정책으로부터의 과도한 이탈을 막기 위해 KL 정규화가 함께 사용된다. 이에 따라 목적함수는 기대 보상을 최대화하면서 참조 정책과의 KL divergence를 제어하는 형태로 정의된다.

$$\max_{\pi_\theta} \mathbb{E}\left[\sum_{t=0}^{H} r(s_t, a_t) - \beta \log \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\text{ref}}(a_t \mid s_t)}\right]$$

3.2. 사후적 기여도 할당 (Hindsight Credit Assignment)

사후적 기여도 할당(Hindsight Credit Assignment, HCA)[9]은 최종 결과를 관찰한 뒤 각 행동의 기여도를 사후적으로 평가하여, 지연되거나 희소한 보상 문제를 완화하는 방법이다. 그 핵심은 총 보상이 주어졌을 때 상태 st에서 행동 at가 선택되었을 조건부 확률인 사후적 분포

$$h(a_t|s_t, z) = P(A_t = a_t | S_t = s_t, Z(\tau) = z)$$

에 있다. RLHF에서는 보상이 보통 종료 상태 sH에서만 주어지므로,

$$h(a_t|s_t, z) = P(A_t = a_t | s_t, r_H)$$

로 쓸 수 있다. 이를 바탕으로 hindsight-based advantage는

$$A^\pi(s_t, a_t) = \mathbb{E}\left[r_H \left(1 - \frac{\pi(a_t|s_t)}{h(a_t|s_t, r_H)}\right)\right]$$

로 정의되며, 이는 최종 보상을 조건으로 했을 때 해당 행동이 현재 정책에 비해 얼마나 더 바람직한지를 측정한다.

4. Hindsight PPO (HPPO)

4.1. 사후적 분포 근사

본 연구의 목표는 표준 PPO에서 사용하는 어드밴티지 함수를 사후적 분포에서 유도된 토큰 수준 어드밴티지로 대체하는 것이다. 그러나 LLM 기반 RLHF에서는 토큰 단위의 거대한 이산 상태 및 행동 공간 때문에 사후적 분포를 직접 계산하기 어렵다. 이를 위해 본 연구는 최종 상태 s_H에 대한 에너지 기반 모델을 도입하여 hindsight distribution을 근사한다. 먼저, 자기회귀 생성에서 s_{t+1}는 s_t와 a_t에 의해 결정되므로, Bayes 규칙을 이용하면 사후적 분포는 다음과 같이 쓸 수 있다.

$$\begin{aligned} P(a_t \mid s_t, r_H) &= \frac{P(r_H, s_t, a_t)}{P(r_H, s_t)} \\ &= \frac{P(r_H|s_t, a_t)P(s_t, a_t)}{P(r_H|s_t)P(s_t)} \\ &= \frac{P(r_H \mid s_{t+1})}{P(r_H \mid s_t)} \cdot \pi(a_t \mid s_t) \end{aligned}$$

위의 경우 가능한 종료 상태의 수가 방대하여 직접 계산이 불가능하기에, 먼저 조건부 확률 $P(r_H \mid s_t)$를 sH을 사용하여 아래와 같은 관계를 도출함.

$$P(r_H \mid s_t) = \sum_{s_H} P(r_H, s_H \mid s_t) = \sum_{s_H} P(r_H \mid s_H) P(s_H \mid s_t).$$

이후 $P(s_H \mid s_t)$를 에너지 기반 모델로 근사한다.

$$P(s_H \mid s_t) = \frac{\exp\big(-E(s_t, s_H)\big)}{Z_{s_t}}, \qquad E(s_t, s_H) = \gamma\, \xi(s_t, s_H),$$

여기서 $\xi(s_t, s_H)$는 중간 상태와 최종 상태 사이의 의미론적 거리이다. 이를 확률 표현식에 대입함으로써, 다음을 얻을 수 있다.

$$P(r_H \mid s_t) = \sum_{s_H} P(r_H \mid s_H) \frac{\exp\big(-\xi(s_t, s_H)\big)}{Z_{s_t}}.$$

위 합은 여전히 계산 비용이 크므로, 실제 LLM이 생성한 최종 상태 sH를 사용하는 plug-in approximation을 적용한다.

$$P(r_H \mid s_t) \approx P(r_H \mid s_H) \frac{\exp\big(-\xi(s_t, s_H)\big)}{Z_{s_t}},$$

$$P(r_H \mid s_{t+1}) \approx P(r_H \mid s_H) \frac{\exp\big(-\xi(s_{t+1}, s_H)\big)}{Z_{s_{t+1}}}.$$

마지막으로 인접한 두 상태에 대해 정규화 상수가 거의 같다고 가정하면, 사후적 분포는 다음과 같이 근사된다.

$$P(a_t \mid s_t, r_H) \approx \frac{\exp\big(-\xi(s_{t+1}, s_H)\big)}{\exp\big(-\xi(s_t, s_H)\big)}\, \pi(a_t \mid s_t).$$

4.2. HPPO를 위한 목적함수 공식화

어드밴티지 함수와 사후적 분포 근사를 통합하여, 어드밴티지 함수는 다음과 같이 공식화된다.

$$A^\pi(s_t, a_t) = \mathbb{E}\left[r_H \left(1 - \frac{\exp(-\xi(s_t, s_H))}{\exp(-\xi(s_{t+1}, s_H))}\right)\right]$$

이 어드밴티지 함수는 표준 PPO 목적 함수 plug-in되어 사용된다.

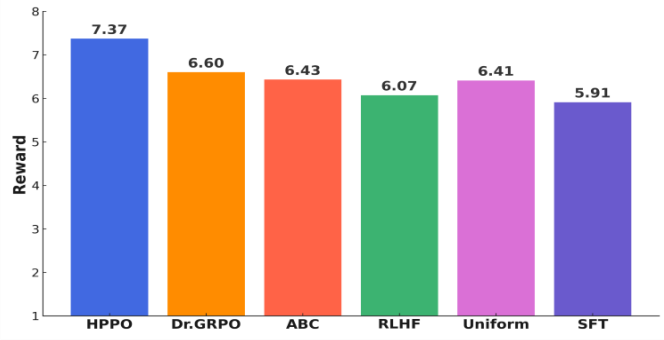
5. 실험결과 (Results)

HPPO의 효과를 검증하기 위해, 본 연구는 정렬 성능과 일반화 성능을 함께 평가한다. 정렬 성능은 보상 모델이 부여한 task-specific reward를 통해 측정하며, 이를 통해 HPPO가 인간 선호 정렬에 미치는 영향을 정량적으로 분석한다. 일반 성능은 널리 사용되는 MT-Bench와 AlpacaEval을 통해 평가한다. 비교 기준으로는 보상 모델의 attention map을 이용해

토큰 수준 보상을 재분배하는 ABC, 문장 수준 피드백만을 사용하는 sentence-level RLHF, hindsight distribution을 균일 분포로 대체한 Uniform Distributionablation, 그리고 강화학습 없이 supervised fine-tuning만 수행한 SFT모델을 사용한다.

5.1. 보상비교

유용성 생성 태스크에서 HPPO는 그림 1와 같이 모든 비교 기준 모델 중 가장 높은 보상을 달성하였다. 이는 사후적 분포를 활용한 토큰 수준 어드밴티지 추정 덕분이다. HPPO의 어드밴티지 함수는 양수와 음수 값을 모두 허용하여 각 토큰의 실제 기여도를 보다 정확히 반영하는 반면, ABC는 비음수 어텐션 가중치에 의존하기 때문에 무관하거나 유해한 토큰에도 양의 신호를 부여할 수 있어 편향이 발생할 수 있다. 반면 RLHF는 희소한 문장 수준 보상에 의존하므로, 짧고 보수적인 응답을 통해 보상을 빠르게 극대화하는 경향이 있어 상대적으로 낮은 평균 보상을 보인다.



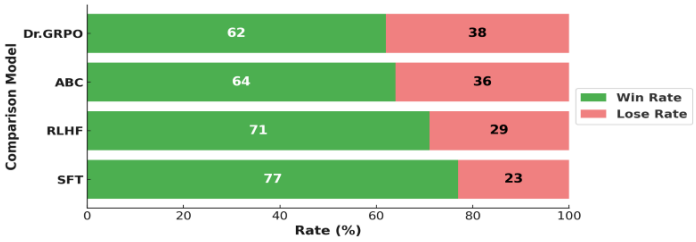
[그림 1. 기준모델별 평균 보상값]

5.2. MT-Bench 및 Alpaca Eval 결과

표 1의 MT-Bench 결과는 HPPO가 Turn 1, Turn 2, 그리고 Overall 전반에서 일관된 성능 향상을 달성했음을 보여준다. 이는 HPPO가 과업 특화 보상 최적화에 그치지 않고 일반 추론 성능 또한 효과적으로 유지함을 시사한다. 한편 AlpacaEval은 인간 선호와의 상관이 더 높은 평가 지표로 알려져 있으며, 그림 2에서 보이듯 HPPO는 이 벤치마크에서도 모든 비교 기준 모델을 능가하였다. 이는 HPPO가 instruction-following 성능과 인간 선호 정렬 측면에서 모두 강한 성능을 보인다는 점을 뒷받침한다.

Model	Turn 1	Turn 2	Overall
HPPO (Ours)	**8.24**	**7.47**	**7.86**
Dr.GRPO	7.35	6.92	7.14
ABC	7.30	6.88	7.09
$RLHF_{sentence}$	6.21	5.33	5.78
SFT	5.16	3.02	4.09

[표 1. 기준모델별 MT-Bench Score]



[그림 2. 기준모델 상호 간 Win-rate (AlpacaEval)]

6. 결론 (Conclusion)

본 연구는 사후적 분포에 기반한 토큰 수준 신용 할당 방법인 HPPO를 제안하였다. HPPO는 출력 임베딩을 활용해 사후적 분포를 근사함으로써 문맥을 반영한 토큰 수준 피드백을 제공하며, 외부 LLM 없이도 효과적인 기여도 추정을 가능하게 한다. 더불어 에너지 기반 모델링을 통해 계산 부담을 줄이면서도 실용적인 근사식을 도출하였다. 실험 결과, HPPO는 인간 선호 정렬 성능을 개선하고 사전학습된 일반화 능력을 유지하면서 기존 방법들을 일관되게 능가하였다.

7. 참고문헌 (Reference)

[1] Yiran Wang et al., Can llms clarify? investigation and enhancement of large language models on argument claim optimization. ACL, page 4066-4077, 2025.

[2] Ouyang et al.,Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35, 2022.

[3] Rafael Rafailov et al., Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems 36: 2023.

[4] Wendi Li et al., Reinforcement learning with token-level feedback for controllable text generation. In Findings of the Association for Computational Linguistics: NAACL, pages 1704-1719, 2024.

[5] Meng Cao et al., Enhancing reinforcement learning with dense rewards from language model critic. EMNLP, pages 9119-9138, 2024.

[6] Eunseop Yoon et al., TLCR: token-level continuous reward for fine-grained reinforcement learning from human feedback. In Findings of the Associa tion for Computational Linguistics, ACL, pages 14969-14981, 2024.

[7] Alex James Chan et al., Dense reward for free in reinforcement learning from human feedback. ICML 2024, Vienna, Austria, July 21-27, 2024.

[8] Zichen Liu et al.,Understanding r1-zero-like training: A critical perspective. CoRR, 2025.

[9] Anna Harutyunyan et al., Hindsight Credit Assignment, Neurips, 2019.