https://doi.org/10.7840/kics.2025.50.5.790

TRPO, PPO, 그리고 DPO를 통한 거대언어모델 강화학습 방법론 동향 연구

김 태 현◆, 박 수 현°

Research on Reinforcement Learning Methodologies for Large Language Models Using TRPO, PPO, and DPO

Taehyun Kim◆, Soohyun Park°

요 약

거대언어모델 훈련에 강화학습을 이용하게 되면서, 거대언어모델을 위한 최적의 강화학습 방법론을 찾아야 할 필요성이 대두되었다. 거대언어모델과 강화학습은 계속해서 새로운 기법을 통해 상호 발전하고 있다. 본 논문에서는 이러한 거대언어모델의 성능 향상을 위한 강화학습 알고리즘의 동향에 대하여 다루었다.

키워드 : 강화학습, 거대언어모델
Key Words : RLHF, LLMs

ABSTRACT

As the utilization of reinforcement learning (RL) in training large language models (LLMs) becomes more prevalent, the necessity to identify optimal RL methodologies tailored for LLMs has emerged. The fields of LLMs and RL are continually evolving through the development of novel techniques that contribute to their mutual advancement. This paper addresses the current trends in reinforcement learning algorithms aimed at enhancing the performance of large language models.

Ⅰ. 서 론

거대언어모델의 등장으로 자연어 처리에서 강화학습의 역할을 확장했으며, 기존보다 더 복잡하고 정교한 자연어 처리 작업이 가능해졌다. 그러나 모델이 사용자 선호를 충분히 반영하지 못하거나 학습 효율성이 떨어지는 문제가 빈번하다[1]. 이러한 문제를 해결하기 위하여 PPO (Proximal Policy Optimization), DPO (Direct Preference Optimization) 등의 강화학습 알고리즘이 개발되었다. 이러한 강화학습과 거대언어모델의 상호작용은 자연어 처리 연구에 앞으로도 상당한 발전을 가져올 것으로 기대된다. 본 논문에서는 거대언어모델의 강화학습 방법론 발전 동향에 대하여 살펴보고자 한다.

Ⅱ. 본 론

TRPO (Trust Region Policy Optimization)는 일관된 학습을 위해 [2]에서 제안된 방법론으로, KL 발산값을 제어하는 접근법을 도입했다. 이 방법론은 이후에 기술할 PPO의 기반이 되었다. TRPO의 아이디어는 이

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 연구결과로 수행되었습니다 (2022-0-01087).
◆ First Author : Sookmyung Women's University Department of Computer Science, puffaaao@sookmyung.ac.kr, 학생회원
° Corresponding Author : Sookmyung Women's University Division of Software, soohyun.park@sookmyung.ac.kr, 정회원
논문번호 : 202410-225-0-SE, Received September 25, 2024; Revised November 30, 2024; Accepted December 18, 2024

전 정책과 새 정책 간의 차이를 의미하는 KL 발산 값이 일정 기준 이하를 유지하도록 하여 정책 업데이트의 기준을 잡는 것이다.

그림 1에서 가로축은 정책 매개변수 θ이고, 세로축은 각 곡선이 가지는 값이다. 녹색 곡선은 θ에 의해 변동되는 Surrogate 함수를, 청색 곡선은 KL 발산을 나타낸다. 그래프에서 적색 영역은 θ가 -0.5에서 0.5 사이의 값을 갖는 범위에서 형성된 신뢰 영역을 나타내는데, TRPO는 θ가 신뢰 영역 내에 있을 때만 정책의 업데이트를 허용하여 안정적인 학습을 가능하게 한다. 그러나 TRPO의 행렬 연산은 막대한 비용과 시간이 소요되어, 거대언어모델에 적용하기 어렵다. 따라서, TRPO의 비실용성을 개선할 수 있는 PPO가 등장하게 되었다. PPO는 TRPO와 마찬가지로 Surrogate 함수를 최대화하는 목적을 가지나, TRPO를 개선하기 위해 Clipping 기법을 도입함으로써 KL 발산 계산 과정을 생략하고 연산 효율성을 향상시킨다[3]. 아래의 식은 [3]에서 제안하는 PPO의 Clipped Surrogate 함수이다.

$$clip(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon)\hat{A}_t \quad (1)$$

위의 수식(1)에서 $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$는 현재 정책이 상태 s_t에서 행동 a_t를 선택할 확률과 이전 정책이 같은 상태 s_t에서 같은 행동 a_t를 선택할 확률비를 나타낸다. ϵ 값은 이 확률비의 범위를 정의하기 위한 하이퍼파라미터이다. 결과적으로, PPO에서는 Clipping 함수를 이용해 $1-\epsilon$과 $1+\epsilon$사이로 확률비의 하한과 상한을 주어, 확률비가 하한보다 작지 않으며 상한보다 크지 않게 제한한다. Clipping 기법은 KL발산을 계산할 필요가 없으므로 정책 업데이트의 연산 과정을 간단히 하여 TRPO의 연산 복잡성 문제를 개선한다. PPO가 연산의 효율과 정책 수렴의 안정을 모두 가져오기에, 거대언어모델과 같은 고차원의 작업에 널리 사용되고 있다.

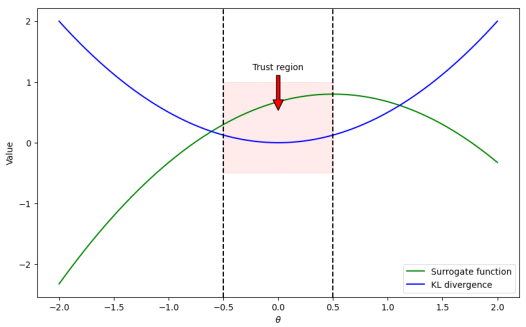

그림 1. Surrogate 함수와 KL 발산으로 형성된 TRPO의 신뢰 영역
Fig. 1. Trust Region of TRPO Represented by Surrogate Function and KL Divergence

DPO기반 학습의 소개에 앞서 RLHF (Reinforcement Learning from Human Feedback)에 대한 개념이 필요하다. RLHF는 강화학습 과정에 인간의 피드백을 활용하여 모델이 사용자 의도에 더 부합하는 결과를 생성하도록 하는 학습법이다. 기존 RLHF는 수집된 인간 피드백을 기반으로 보상 모델을 설계하고, 이를 통해 정책을 평가하는 방식으로 작동한다. 이때 PPO를 사용하는 기존의 RLHF 알고리즘은 인간의 피드백으로부터 보상 모델을 구체화하는 복잡한 과정이 필수적이라는 한계를 갖는다[4]. 하이퍼파라미터에 민감하여 하이퍼파라미터 설정에도 시간과 비용 부담이 크다는 것 또한 단점이다. 더불어, 인간 피드백을 수집하는 과정 또한 큰 부담을 가져온다.

따라서, 하이퍼파라미터 설정에 덜 민감하며 복잡한 보상 모델을 제거하고, 데이터 수집 요구를 최소화한 방법론인 DPO가 [4]에서 제시되었다. 다음 그림 2에서 DPO가 어떻게 기존 RLHF의 문제점들을 개선하는지 살펴볼 수 있다. 그림 2는 사용자가 '영화 리뷰를 작성하라'는 입력을 준 상황을 나타낸다. 입력이 주어지면, 모델은 입력 데이터를 읽고 여러 가지의 리뷰를 생성한다. 그 후, 사용자의 선호 정도를 생성된 여러 개의 리뷰 중 사용자의 피드백을 기반으로 데이터에 라벨링하여 선호 데이터를 구성한다. DPO는 이 사용자 선호 데이터를 직접 거대언어모델 학습에 이용하여, 모델이 사용자의 선호를 반영할 수 있도록 설계되었다. 이 과정은 선호 데이터를 학습한 모델로 새로운 선호 데이터 입력을 처리하고, 반복적으로 개선하는 구조로 이루어진다. DPO의 핵심은 명시적인 보상 모델에 의존하지 않고, 직접 인간의 피드백 데이터로부터 모델을 평가하고 최적화하여 학습하는 방식을 사용하는 것에 있다. 선호

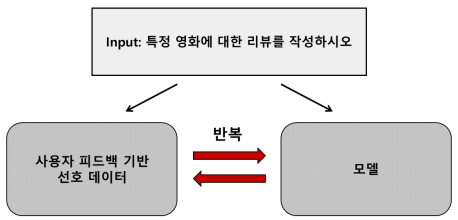

그림 2. 사용자의 피드백을 기반으로 선호 데이터를 라벨링하고 이를 활용해 DPO가 모델을 반복적으로 개선하는 과정
Fig. 2. The process of labeling preference data based on user feedback and using it to iteratively improve the model through DPO

데이터를 곧바로 학습에 이용하기에 하이퍼파라미터 설정의 중요성이 줄어들고, 선호 데이터가 보상 모델 역할을 하게 되며, 반복적인 구조로 데이터의 재사용이 이루어져 피드백 데이터 수집의 부담이 줄어든다. 결과적으로 DPO는 학습의 효율성을 높이고, 거대언어모델이 인간의 선호에 알맞은 출력을 생성하도록 돕는다.

Ⅲ. 결 론

본 논문에서는 거대언어모델을 위한 강화학습 방법론을 소개하였다. TRPO는 KL 발산 값을 기준으로 안정적인 학습을 시도했다는 점에서 의의를 가지지만, 연산 복잡성 문제로 인해 PPO의 Clipping 기법이 등장했다. 이어서, DPO의 학습 방식과 장점을 살펴보았다. DPO는 RLHF의 한계를 보상 모델 설계 과정을 제거하고 데이터 수집 부담을 완화함으로써 개선했지만, 사용자 피드백 데이터 품질에 의존하는 경향이 있고 과적합 가능성 등의 단점이 존재해 이를 개선할 수 있는 새로운 강화학습 방법론이 필요하다. 최근 높은 성능의 거대언어모델이 상업적 서비스에 널리 사용되고 있는 반면, 모델의 구체적인 학습 프레임워크에 대한 공개가 희박하여 정보의 제한이 존재한다. 이는 거대언어모델 강화학습 방법론 연구를 확장하기 어려운 환경을 조성하고 있다[5]. 따라서 연구자들이 활용 가능한 데이터와 알고리즘 접근성을 확대하여 거대언어모델 강화학습 방법론의 연구를 활성화하는 것이 필요하다. 이러한 노력은 인공지능 시대의 거대언어모델이 보다 효율적이면서 신뢰할 수 있는 방향으로 진화하는 데 중요한 역할을 할 것이다.

References

[1] S. Y. Kim, J. B. Shin, H. G. Yoon, J. S. Lee, and H. J. Cho, "Technology trends of large language models in the age of generative AI," in *Proc. KIISE*, vol. 41, no. 11, pp. 25-33, Nov. 2023.

[2] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," in *Proc. ICML*, pp. 1889-1897, Lille, France, 2015.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *OpenAI*, Jul. 2017.

[4] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Proc. Conf. NeurIPS*, pp. 53728-53741, New Orleans, LA, USA, Dec. 2023.

[5] J. Kim, A. I. Danielle, and H. Lim, "A study about efficient method for training the reward model in RLHF," in *Proc. HCLT*, pp. 245-250, Jeju, Korea, Oct. 2023.

김 태 현 (Taehyun Kim)

2022년 3월~현재 : 숙명여자대학교 컴퓨터과학과 학사과정

<관심분야> 강화학습, 거대언어모델

[ORCID:0009-0004-7972-6863]

박 수 현 (Soohyun Park)

2019년 2월 : 중앙대학교 소프트웨어대학 컴퓨터공학과 졸업 (공학사)

2023년 8월 : 고려대학교 공과대학 전기전자공학과 졸업 (공학박사)

2023년 9월~현재 : 숙명여자대학교 공과대학 소프트웨어학부 조교수

<관심분야> Deep Learning Theory, Network /Mobility Applications, Quantum Machine Learning, AI-based Autonomous Control

[ORCID:0000-0002-6556-9746]