TRPO, PPO, 그리고 DPO 를 통한 거대언어모델 강화학습 방법론 동향 연구

김태현, 박수현
숙명여자대학교

puffaaao@sookmyung.ac.kr, soohyun.park@sookmyung.ac.kr, *joongheon@korea.ac.kr

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

Taehyun Kim, Soohyun Park
Sookmyung Women's Univ.

요 약

기존의 간단한 강화학습 시나리오들이 아닌 복잡한 거대언어모델에 강화학습을 적용하게 되면서, 거대언어모델을 위한 최적의 강화학습 방법론을 찾아야 할 필요성이 대두되었다. 거대언어모델과 강화학습의 유기적 관계는 계속해서 새로운 기법을 통해 발전되고 있다. 본 논문에서는 이러한 거대언어모델의 성능 향상을 위한 강화학습 알고리즘의 동향에 대해 다루었다.

Ⅰ. 서 론

강화학습은 에이전트가 환경과 상호작용하며 스스로 훈련하는 머신러닝 방법론이다. 거대언어모델(LLM: Large Language Models)의 등장 전에도 강화학습은 자연어 처리 분야에 사용되었으나, 강화학습에 자연어 처리를 적용할 때 적절한 보상 모델을 선택하는 것은 복잡한 과제로 인식되었다. 강화학습이 자연어 처리 분야에 효과적으로 적용되기 위해서는 정밀한 보상 체계가 필수적이기 때문이다. 그렇기에 거대언어모델 이전의 강화학습 기법은 주로 간단한 Gaming 시나리오와 같은 비교적 덜 복잡한 문제 해결에 집중되었다. 거대언어모델의 등장으로 최근 몇 년간 강화학습의 역할은 자연어 처리 분야에서 크게 확장되었으며, 거대언어모델은 강화학습 기법으로부터 도움을 받아 더욱 복잡한 자연어 처리 작업을 처리할 수 있게 되었다. 그러나 GPT-3와 같은 대규모 사전학습 언어모델은 사용자의 선호와 어긋나는 유해한 답변이나 환각 현상을 일으키는 문제를 겪었다[1]. 이를 해결하기 위한 PPO (Proximal Policy Optimization), DPO (Direct Preference Optimization) 등의 강화학습 알고리즘의 등장이 거대언어모델의 자연어 이해 및 생성 역량을 강화하고 있다. 이러한 강화학습과 거대언어모델의 상호작용은 자연어 처리 연구에 있어 앞으로도 상당한 발전을 가져올 것으로 기대된다. 따라서, 본 논문에서는 거대언어모델의 강화학습 방법론 발전 동향에 대하여 살펴보고자 한다.

Ⅱ. 본론

TRPO (Trust Region Policy Optimization)는 일관된 학습을 유지하기 위해 등장한 방법론이자 후에 기술할 PPO 알고리즘의 전신이다. TRPO 의 아이디어는 이전 정책과 새 정책 간의 차이를 의미하는 KL 발산 값이 일정 기준 이하를 유지하도록 하여 정책의 업데이트 기준을 잡는 것이다[2]. 다시 말해, Surrogate 함수의 KL 발산을 제한해 정책의 급격한 업데이트를 제한한다. 그림 1 로 KL 발산 값에 따른 신뢰 영역을 시각적으로 확인할 수 있다.

그림 1 에서 가로축은 정책의 매개변수를 나타내고, 세로축은 두 개의 곡선이 가지는 값이다. 녹색 곡선은 정책

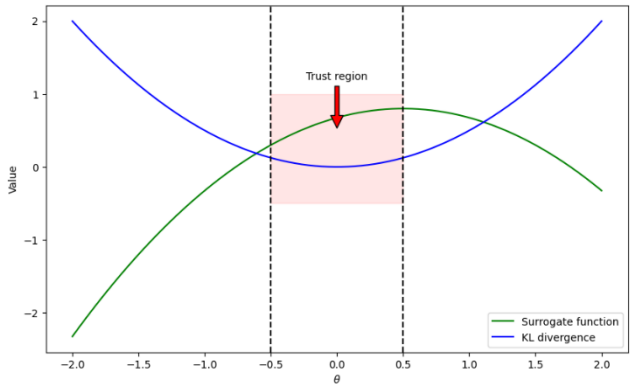

그림 2 TRPO 의 신뢰 영역

매개변수 θ에 의해 변동되는 Surrogate 함수이며, 청색 곡선은 KL 발산을 표현한다. 그래프에서 적색으로 강조된 부분이 형성된 신뢰 영역을 나타내는데, TRPO 는 θ가 -0.5 에서 0.5 사이의 신뢰 영역 내에 있을 때만 정책 업데이트를 허용해 안정적인 학습을 가능하게 한다. 그러나, TRPO 의 한계는 연산의 복잡성에 있다. TRPO 의 행렬 연산은 그 시간과 비용이 매우 많이 들어, 거대언어모델에 적용하기 어렵다. 따라서, TRPO 의 비실용성을 개선할 수 있는 PPO 가 등장하게 되었다.

PPO 또한 Surrogate 함수를 최대화하는 목적을 가지지만, TRPO 와 다르게 계산의 복잡도를 줄이기 위하여 Clipping 기법을 사용한다[3]. 아래의 식은 PPO 의 Clipped Surrogate 함수이다.

$$\text{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon\right)\widehat{A_t}$$

위의 수식에서 $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$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. 확률비 $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$가 너무 작은 경우에는 local optima 에 빠지거나 학습이 더딜 수 있고, 너무 큰 경우에는 과적합이 생길 수 있으며 학습이 쉽게 불안정해지기에, 확률비를 제한하는 정도를 결정하는 ϵ 값은 Clipping 기법에서 중요한 역할을 한다. 결과적으로, PPO 에서는 Clipping 함수를 이용해 $1-\epsilon$과 $1+\epsilon$사이로 확률비 $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$의 하한과 상한을 주어, 확률비가 하한보다 작지 않으며 상한보다 크지 않게 제한한다. Clipping 기법은 정책의 업데이트를 간단히 하고 안정적 수렴을 보장하여 TRPO 의 연산 복잡성 문제를 개선한다. PPO 가 연산의 효율과 정책 수렴의 안정을 모두 가져오기에, 거대언어모델과 같은 고차원의 데이터를 다루는 작업에 널리 사용되고 있다.

DPO 기반의 학습을 소개하기에 앞서 RLHF (Reinforcement Learning from Human Feedback)에 대한 개념이 필요하다. RLHF 는 강화학습 프로세스에 인간의 피드백을 추가하여 인간의 추구하는 방향에 모델이 더욱 가까워지도록 하는 학습법이다. 일반적인 RLHF 는 인간의 피드백을 수집하고, 그를 보상 모델로 가공하여 정책을 평가하게 한다. 이때 RLHF 는 PPO 를 포함한 다양한 강화학습 알고리즘을 선택하여 모델을 최적화할 수 있다. PPO 를 사용하여 RLHF 학습 수행 시 안정성 있는 정책 업데이트가 가능하다는 장점이 있으나, PPO 를 포함한 기존의 RLHF 알고리즘은 공통적인 문제를 갖는다. 인간의 피드백을 반영한 데이터 셋으로부터 보상 모델을 체계화하는 복잡한 과정이 필수적인 것이 원인이다[4]. 하이퍼파라미터들에 민감하여 하이퍼파라미터 설정 과정에도 시간과 비용이 부담이 크다는 것 또한 단점이다. 예를 들어, 학습률에 민감하여 정밀한 학습률 설정 없이는 수렴이 과도하게 지연되거나, 오버슈팅 가능성이 높은 경향이 있다. 더불어, 인간 피드백을 수집하는 데이터 샘플링 과정 또한 큰 부담을 가져온다.

따라서, 하이퍼파라미터 설정에 덜 민감하며 복잡한 보상 모델을 제거하고, 데이터 샘플링 요구를 최소화한 DPO 가 기존 RLHF 의 대체제로 나타나게 되었다. 다음 그림 2 에서 DPO 가 어떻게 기존 RLHF 의 문제점들을 개선하는지 살펴볼 수 있다. 그림 2 는 특정 문장을 생성하도록 하는

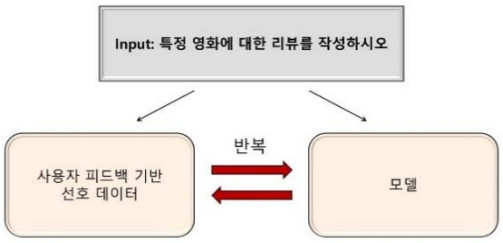

그림 1 DPO 적용 예시

사용자의 입력을 받은 상황을 나타낸다. 사용자가 '영화 리뷰를 작성하라'는 입력을 주면, 모델이 입력 데이터를 읽고 여러 가지의 리뷰를 생성한다. 그 후, 생성된 여러 개의 리뷰 중 사용자의 피드백을 기반으로 사용자의 선호 정도를 데이터에 라벨링 해 선호 데이터를 구성한다. DPO 는 이 선호 데이터를 직접 모델 학습에 이용해 사용자의 선호를 반영하게 하고, 선호 데이터를 학습한 모델로 입력 데이터를 처리하는 반복 구조를 가져, 모델이 지속적으로 개선된다.

DPO 의 핵심은 명시적인 보상 모델 학습에 의존하지 않고, 인간의 피드백 데이터로부터 모델을 평가하고 최적화해 곧바로 거대언어모델을 학습하는 방식을 사용하는 것에 있다. 선호 데이터를 직접 학습에 이용하기에 하이퍼파라미터 설정의 중요성이 줄어들고, 선호 데이터가 보상 모델 역할을 하게 되며, 반복적인 학습 구조로 데이터의 재사용이 이루어져 피드백 데이터 수집의 부담이 줄어든다. 결과적으로 DPO 는 학습의 효율성을 높이고, 인간의 선호에 알맞은 출력을 생성하도록 돕는다.

Ⅲ. 결론

본 논문에서는 거대언어모델을 위한 강화학습 방법론에 대해 소개하였다. TRPO 는 KL 발산 값으로 정책 업데이트의 기준을 잡아 안정적인 학습을 시도했다는 의의가 있다. TRPO 는 연산이 과도하게 복잡하다는 단점이 있었으나, 이는 PPO 의 Clipping 기법으로 개선할 수 있었다. 이어서, PPO 를 이용하는 기존 RLHF 의 보상 모델 학습의 번거로움을 비롯한 단점들과 그 개선책인 DPO 에 대해서도 알아보았다. 본 연구에서는 DPO 까지의 강화학습 방법론 탐색의 동향을 정리하였으나, DPO 또한 과적합 문제 등의 한계가 존재한다. 이를 보완하고 강화학습의 최적 방법론을 모색하기 위해, KTO (Kahneman-Tversky Optimization), ORPO (Odds Ratio Preference Optimization), ITO (Identity Preference Optimization) 등의 강화학습 방식들이 연구되고 있다. 인공지능 시대 거대언어모델의 진화에 발돋움이 될 강화학습 기법들의 중요성은 더욱 커질 것이며, 이에 대한 지속적인 관심과 연구가 필요하다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 연구결과로 수행되었음 (2022-0-01087)

참 고 문 헌

[1] S. Yeon Kim, J. Bo Shin, H. Gu Yoon, J. Sik Lee, and H. Jik Cho, "Technology Trends of Large Language Models in the Age of Generative AI," J. Korean Inst. Inf. Sci. Eng. (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," arXiv:1502.05477, University of California, Berkeley, Department of Electrical Engineering and Computer Sciences, Feb. 2015, last revised Apr. 2017.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv:1707.06347, OpenAI, Jul. 2017, last revised Aug. 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," arXiv:2305.18290, 2023.