인간 선호를 더 정확히 구분하기 위한 Quantile Regression 기반 보상 모델링

김민성 [1], 홍민식 [1], 김동재 [2]
[1] 단국대학교 소프트웨어학과, [2] 단국대학교 인공지능융합학과
{kms0509, minsik.hong, dongjaekim}@dankook.ac.kr

Quantile Regression-Based Reward Modeling for Improved Discriminating Human Preferences in Language

Min Seong Kim[1], Min Sik Hong[1], Dong Jae Kim[2]
[1]Department of Software, [2]Department of AI-based Convergence
{kms0509, minsik.hong, dongjaekim}@dankook.ac.kr

요 약

최근 대규모 언어 모델(Large Language Model, LLM)의 발전은 인간 피드백을 활용한 강화학습(Reinforcement Learning From Human Feedback, RLHF)을 통해 언어 모델의 응답을 인간 선호도에 정렬하는 방향으로 이루어지고 있다. 기존 RLHF 에서는 보상 값을 단일 스칼라 형태로 예측하는 방식을 주로 사용해왔으나 이러한 방식은 보상의 불확실성이나 인간 선호의 다양성을 충분히 반영하지 못하는 한계가 있다. 본 논문에서는 기존 RLHF 에서 사용되는 스칼라 보상 예측 방식의 한계를 극복하기 위해 Quantile Regression 기반의 보상 분포 예측 구조를 도입하였다. 인간 선호 데이터를 기반으로 학습되는 본 보상 모델은 응답에 대한 단일 보상 값이 아닌 분포 형태의 보상 출력을 예측함으로써 보다 세밀한 인간 선호 정보를 포착할 수 있다. 실험에는 HH(Harmlessness-Helpfulness) 응답 비교 데이터셋을 활용하였으며 기존 스칼라 보상 방식과 제안된 분포 방식 간의 정렬 성능을 비교한 결과, 제안된 방식이 인간 선호 정렬 측면에서 더 선호 구분 가능성이 있음을 확인하였다.

1. 서 론

[1]대규모 언어 모델(Large Language Models, LLM)의 발전은 인공지능이 인간의 의도와 가치에 부합하도록 정렬되는 문제를 중심으로 급격하게 발전해왔다[1]. 초기 LLM 은 대규모 사전학습만으로는 사용자 지시를 정확히 반영하거나 안전하고 유연한 출력을 생성하는 과정에 있어 한계를 보였다.

이러한 한계를 극복하기 위해 인간 피드백 기반 강화학습(Reinforcement Learning from Human Feedback, RLHF)이 LLM 에 도입되었다[2]. RLHF 는 LLM 을 인간 선호에 맞추어 미세 조정하는 기술로 여러 LLM 들의 핵심 학습 과정으로 자리 잡았으며 대표적으로 OpenAI 에서는 GPT-3 기반 모델에 RLHF 를 적용하여 InstructGPT 를 개발하였다[3]. 이처럼 RLHF 는 최근 LLM 개발에서 정렬을 위한 표준적 절차로 정착하였다.

RLHF 의 동작 방식은 인간 선호 데이터를 기반으로 보상 모델을 학습하고 이를 이용해 정책 모델을 강화학습으로 최적화하는 두 단계로 수행된다[4]. 이 중 보상 모델은 인간의 의도를 근사하는 중요한 방식이며 보상 모델의 품질이 최종 LLM 성능에 직접적인 영향을 미친다. 하지만 현재까지의 보상 모델은 대부분 스칼라 값 하나를 출력하는 방식에 의존하고 있다. 이는 출력 품질에 대한 다양한 불확실성이나 편차를 충분히 포착하지 못하며 다양한 기준을 반영하기에는 한계가 존재한다. 본 논문에서는 이를 극복하기 위해 Quantile Regression(QR) 기반 분포 강화학습 구조를 보상 학습에 적용하여 스칼라 보상 대신 다수의 분위 수(quantile)를 통해 보상을 예측하는 방식을 제안한다[5].

2. 배경

2.1 RLHF 를 통한 LLM 학습 방법

LLM 은 사전학습을 통해 방대한 범용 지식을 습득하지만, 사용자 의도에 부합하는 응답을 생성하기 위해서는 정렬 과정이 필요하다. RLHF 는 이러한 정렬을 달성하기 위한 대표적인 방법은 일반적으로 다음 세 단계로 구성된다.

첫 번째 단계는 지도학습으로, 초기 LLM 을 고품질 프롬프트-응답 데이터셋에 대해 미세 조정하여 기본적인 프롬프트 대응 능력을 부여한다. 두 번째 단계는 보상 모델 학습으로 인간 평가자가 여러 응답 후보에 대해 선호도를 비교한 데이터를 수집하고 이를 바탕으로 인간

[1] 본 연구는 과학기술정통신부의 재원으로 한국연구재단(RS-2024-00348149) 및 정보통신기획평가원의 SW 중심대학사업(2024-0-00035), 학석사연계 ICT 핵심인재양성사업(IITP-2023-RS-2023-00259867), 대학 ICT 연구센터사업의(IITP-2024-RS-2024-00437102)의 지원을 받아 수행된 연구임.

선호를 근사하는 보상 모델을 학습한다. 마지막 세 번째 단계에서는 학습된 보상 모델을 활용하여 정책 모델을 강화학습 방식으로 최적화한다.

이 중 보상 모델 학습은 RLHF 의 핵심적인 역할을 수행한다. 보상 모델은 주어진 입력 프롬프트 $\boldsymbol{x}$와 이에 대응하는 응답 $\boldsymbol{y}$ 에 대해 인간 평가자의 선호도를 근사하는 점수 $\boldsymbol{r_\phi(x,y)}$를 출력한다. 보상 모델을 학습하기 위해 Bradley-Terry 모델을 기반으로 한 pairwise loss 가 사용되며, 이때의 확률은 다음과 같이 정의된다[6].

$$\boldsymbol{L(\phi) = -log\,\sigma(r_\phi(x,y_{chosen}) - r_\phi(x,y_{rejected}))}\ . \quad (1)$$

이러한 방식은 구조가 간단하고 효과적이지만 각 응답에 대해 하나의 스칼라 보상만을 예측하기 때문에 출력 품질에 대한 다양한 불확실성이나 인간 선호의 복잡한 분포를 충분히 반영하기 어렵다는 한계가 존재한다.

2.2 분포 강화 학습과 Quantile Regression

기존 강화학습은 보상이 스칼라로 출력되지만, 실제 환경에서는 보상이 다양한 원인에 의해 분산되거나 다중 양상으로 나타나는 경우가 빈번하다. 이를 반영하기 위해 분포 강화학습(Distributional Reinforcement Learning)이 제안되었다[7]. 분포 강화학습은 기대 값이 아닌 보상 전체 분포를 학습 대상으로 지정하여 장기 보상의 불확실성, 다양성 등을 정량적으로 포착할 수 있다.

분포 강화학습 기법 중 하나인 QR-DQN(Quantile Regression Deep Q-Network)은 전체 보상 분포를 N 개의 균등한 분위 수(quantile)로 나눈 뒤, 각 고정된 확률 레벨에 해당하는 분포의 위치를 직접 회귀함으로써 보상 분포를 정밀하게 근사한다. 이 방식은 Wasserstein metrics 기반의 최적화가 가능하고 단일 스칼라 값을 예측하는 기존 방식과 달리 분포의 형태를 학습할 수 있어 학습 안정성과 표현력 측면에서 우수한 성능을 보인다.

3. 연구 방법

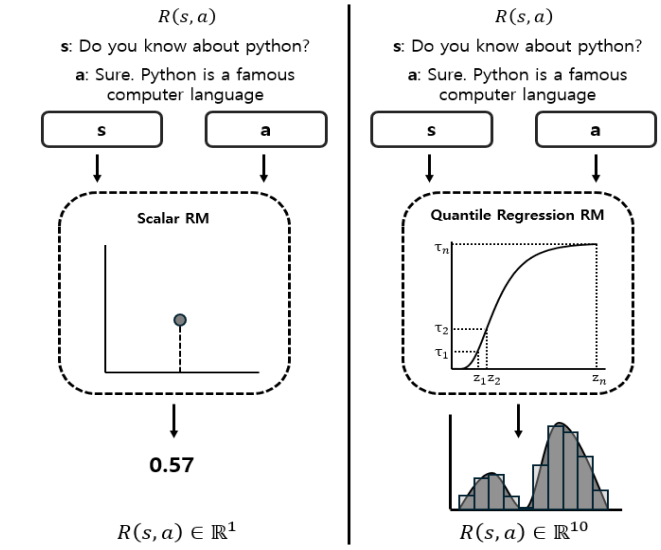

그림 1. 스칼라 보상 모델과 QR 보상 모델

3.1 연구 동기

본 논문에서는 QR-DQN 에서 제안된 QR 기반의 분포 예측 구조를 RLHF 기반 LLM 의 보상 모델 학습 과정에 도입한다. 기존 RLHF 기반 LLM 에서는 보상 모델이 각 응답에 대해 단일 스칼라 값을 출력하는 구조를 사용하였으나 본 논문에서는 이를 다수의 분위 수 출력을 가지는 분포 형태로 확장함으로써, 인간 선호의 불확실성과 다양성을 세밀하게 포착하고자 한다(그림 1).

3.2 실험 설정

기존의 스칼라 보상 모델과 본 논문에서 제안하는 QR 기반 보상 분포 모델을 비교하여, 분포 기반 접근이 인간 선호의 정렬 측면에서 가지는 차이를 분석한다. 모델 구조는 MOSS-RLHF 의 보상 모델 학습 과정을 기반으로 구현되었으며, 오픈 소스인 GPT-2 모델을 사용하였다[8]. 스칼라 보상 모델은 출력 응답에 대해 하나의 스칼라 값을 예측하며, QR 모델은 동일한 구조에서 출력층을 수정하여 10 개의 분위 수 값을 출력하도록 변경하였다. 학습에 사용된 MOSS-RLHF 의 최적화 과정은 데이터쌍 비교(pairwise)에 따른 손실(log-sigmoid, 수식(1)) 계산 및 quantile regression 을 위한 Huber 손실을 결합한 형태이다[7]. 사용된 데이터셋은 Anthropic 에서 공개한 HH(Harmlessness-Helpfulness) 응답 비교 데이터이며, chosen/rejected 응답 쌍을 기반으로 보상 모델을 학습하였다[9]. 평가 또한 동일한 데이터셋의 검증 데이터(validation split)를 기반으로 수행하였다.

4. 실험결과

4.1 각 보상 모델 방식의 보상 분포 비교

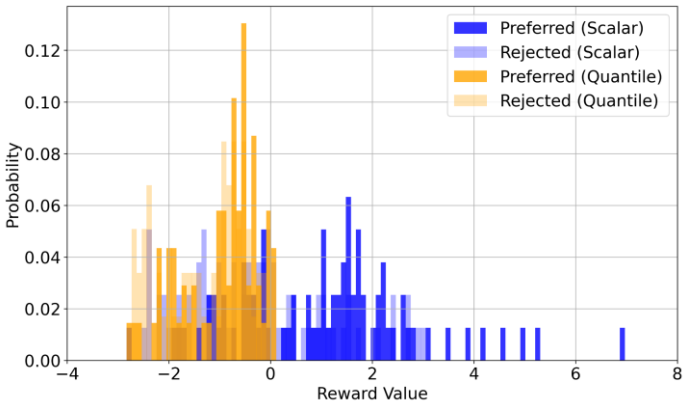

그림 2. Helpful 워크로드의 각 보상 모델 별 분포

우선 Helpful 워크로드에 대해 스칼라 방식과 QR 방식의 보상 모델을 각각 적용한 뒤 산출된 보상 값의 분포를 시각화하였다(그림 2). 스칼라 보상 모델의 경우 선호 응답에 대한 보상 분포가 전체적으로 우측으로 이동해 있는 반면, QR 보상 모델에서는 선호 응답과 비

선호 응답의 보상 분포가 상대적으로 좌측에 집중되어 서로 다른 양상을 보인다.

표 1. 보상 모델 별 각 워크로드의 선호-비 선호 문장 간 KL Divergence 값

Workload	스칼라 보상 모델	QR 보상 모델
Helpful	4.914	6.779
Harmless	6.833	7.291

더 자세한 분석을 위해 스칼라 보상 모델과 QR 보상 모델이 생성한 보상 분포로부터 선호 응답과 비 선호 응답 간의 KL divergence 를 측정하였다(표 1). Helpful 와 Harmless 워크로드 모두 QR 보상 모델이 스칼라 보상 모델보다 더 큰 KL divergence 값을 기록하였으며 특히 Helpful 워크로드에서는 QR 보상 모델이 약 38% 더 높은 차이를 보였다. 이는 QR 보상 모델이 두 응답 간의 차이를 더 뚜렷하게 구분할 수 있으며 인간 선호에 대한 판별력이 더 높다는 점을 시사한다.

4.2 높은 분위수와 낮은 분위수의 분석

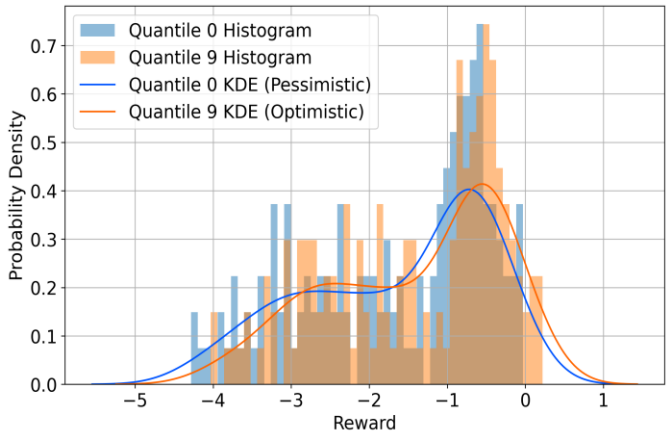

그림 3. Helpful 에서 quantile 0, 9 번에 대한 보상분포

분위 수에 따라 분포의 상위 부분에 해당하는, 낙관적(optimistic) 분포를 나타내거나 하위 부분에 해당하는 비관적(pessimistic) 분포를 나타낸다[7]. 이를 확인하기 위해 분위 수 0(τ =0.05)과 분위 수 9(τ=0.95)에서의 전체 보상 분포를 히스토그램과 커널 밀도 추정(Kernel Density Estimation, KDE) 곡선으로 비교하였다(그림 3). 분위 수 0 에서는 보상 값이 음수 영역에 보다 집중되어 있어 전반적으로 낮은 보상이 할당되는 경향을 보이며, 이는 비관적 관점에서의 보상 할당 특성을 반영한다. 반면, 분위 수 9 에서는 분포가 상대적으로 오른쪽으로 이동하며 높은 보상 값이 더 많이 나타나 낙관적 관점에서의 보상 경향이 두드러진다. 이러한 차이는 분위 수 기반 보상 모델이 동일한 입력에 대해서도 관점(quantile)에 따라 보상을 상이하게 할당함으로써, 보상의 불확실성과 다양성을 효과적으로 포착하고 있음을 보여준다.

5. 결론 및 향후 연구

본 논문에서는 기존 RLHF 보상 모델의 스칼라 방식이 가진 한계를 극복하기 위해 QR 기반 보상 모델을 도입하여 인간 선호를 더 세밀하게 반영할 수 있는 구조를 제안하였다. 제안한 QR 보상 모델은 각 응답에 대해 다수의 분위 수 값을 출력함으로써 출력 분포 전체를 학습 대상으로 삼아 더 세밀하고 표현력 높은 보상 예측이 가능함을 보였다. 실험에서는 HH 데이터셋을 활용하여 스칼라 보상 모델과 QR 보상 모델을 비교하였다. 그 결과, 보상 분포 그래프와 선호-비 선호 문장 간 KL divergence 분석을 통해 QR 보상 모델이 인간 선호를 보다 명확히 구분할 수 있음을 확인하였다. 향후 연구에서는 특정 분위 수 구간이 갖는 의미를 보다 정량적으로 분석할 것이고 QR 보상 모델을 이용한 실제 정책 최적화 단계까지 연계하여 최종 응답의 품질을 효과적으로 개선할 것이다.

6. 참고 문헌

[1] Wang, Zhichao, et al. "A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more." *arXiv preprint arXiv:2407.16216* (2024).

[2] Kirk, Robert, et al. "Understanding the effects of rlhf on llm generalisation and diversity." *arXiv preprint arXiv:2310.06452* (2023).

[3] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022): 27730-27744.

[4] Zheng, Rui, et al. "Secrets of rlhf in large language models part i: Ppo." *arXiv preprint arXiv:2307.04964* (2023).

[5] Bellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." International conference on machine learning. PMLR, 2017.

[6] Bradley, Ralph Allan, and Milton E. Terry. "Rank analysis of incomplete block designs: I. The method of paired comparisons." *Biometrika* 39.3/4 (1952): 324-345.

[7] Dabney, Will, et al. "Distributional reinforcement learning with quantile regression." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.

[8] Wang, Binghai, et al. "Secrets of rlhf in large language models part ii: Reward modeling." *arXiv preprint arXiv:2401.06080* (2024).

[9] Bai, Yuntao, et al. "Training a helpful and harmless assistant with reinforcement learning from human feedback." *arXiv preprint arXiv:2204.05862* (2022).