Truncated Quantile Critics 를 이용한 이동로봇 자율주행 경로 계획

배상욱, 은승우, 한동석*, 김학민, 오재윤
경북대학교, 경북대학교, *경북대학교, 테크로보틱스에어리어, 테크로보틱스에어리어

sangukbae99@gmail.com, sweunwave@knu.ac.kr, *dshan@knu.ac.kr,

kimhakmin@traslam.com, ohjaeyoon@traslam.com

Autonomous Path Planning for Mobile Robots Using Truncated Quantile Critics

Bae Sang Uk, Eun Seung Woo,Han Dong Seog*, Kim Hak Min, Oh Jae Yoon
School of Electronic and Electrical Engineering, Kyungpook National Univ.
Tech Robotics Area

요 약

본 논문은 360° LaserScan 을 직접 상태로 사용하는 Truncated Quantile Critics(TQC) 기반 자율주행 로컬 플래너를 제안한다. 상위 분위수 절단과 자동 온도 조절을 결합하여 과대추정을 억제하고 탐험–이용 균형을 자동화했으며, 우선순위 리플레이·폴리악 타깃·체크포인트 운영으로 표본 효율과 정책 안전성을 강화하였다. 그 결과 평균 Q 의 안정화와 Q_max 의 확장, 낮은 Critic 손실, 그리고 α 의 급격한 감소가 확인되며, 다양한 환경에서도 안정적 경로 계획이 가능함을 보였다.
.

Ⅰ. 서 론

자율주행 로봇의 경로 계획은 인지, 의사결정, 제어를 하나로 통합해야 하는 고난도 문제다. 특히 실내·창고·연구시설처럼 동적 장애물이 많고 구조가 비정형적인 공간에서는 라이다 반사 특성 변화와 센서 노이즈로 인해 전통적 전역·지역 계획만으로 예기치 못한 상황에 기민하게 대응하기 어렵다. 이 한계를 보완하기 위해 센서 입력에서 곧바로 연속 제어 명령을 생성하는 심층강화학습(DRL) 접근이 주목받고 있다.

본 연구는 분포형 가치학습을 채택한 최대엔트로피 강화학습 계열 기법인 Truncated Quantile Critics(TQC)를 자율주행 로봇의 지역 경로 계획에 적용한다. TQC 는 가치함수를 하나의 스칼라가 아니라 분위수 집합으로 추정해 불확실성을 표현하고, 상위 분위수 일부를 의도적으로 제거해 과대추정을 줄인다. 여기에 자동 온도 조절을 결합해 탐험과 이용의 균형을 학습 과정에서 스스로 맞춘다. 그 결과 정책은 기대보상만이 아니라 위험까지 고려하며, 센서 노이즈와 환경의 비정상성에도 더 견고하게 반응한다.

센싱 파이프라인은 Ouster 라이다를 사용하며, 포인트클라우드를 LaserScan 으로 변환해 얻은 360 도 거리 정보를 시간 동기화된 단일 관측 벡터로 구성하여 상태로 직접 투입한다. 행동은 가우시안 정책 출력을 쌍곡탄젠트로 스쿼시한 뒤 실제 로봇의 범위로 스케일링하여 선속과 각속을 연속 값으로 내며, 이러한 센서 직결 구조는 설계 의존도를 낮추고 다양한 환경으로의 이전 가능성을 높인다.

학습 안정성을 위해 다중 크리틱과 다중 분위수로 값함수를 근사하고, 목표 네트워크는 폴리악 평균으로 점진 갱신한다. 리플레이 버퍼는 우선순위 재생을 사용해 TD 오차가 큰 전이를 더 자주 샘플링함으로써 표본 효율을 높인다. 또한 실험 중 에피소드 최소 보상을 추적하여 성능 저하 신호가 나타나면 조기 배치 학습을 수행하고, 일정 간격으로 현재 정책을 체크포인트 액터에 스냅샷해 평가 단계에서 더 안전한 정책을 선택하고 유지할 수 있게 했다

Ⅱ. 본론

본 연구의 모델은 최대엔트로피 원리를 따르는 TQC 로, 가우시안 정책을 학습하는 액터와 분포형 가치함수를 추정하는 다중 크리틱으로 이뤄진다. 상태는 Ouster 라이다를 LaserScan 360° 거리 벡터로 변환해 직접 입력하고, 액터는 평균·로그분산을 예측한 뒤 재매개변수화로 샘플링하고 tanh 스쿼시(야코비안 보정 포함)로 선속·각속을 출력한다. 크리틱은 분위수 집합으로 Q 분포를 근사하며, TQC 는 다음 상태 타깃 분위수를 정렬해 상위 분위수 일부를 절단함으로써 과대추정을 억제하고 학습을 안정화한다. 절단 폭은 “크리틱 수×분위수 수”에서 크리틱별 버릴 분위수 개수를 뺀 양으로 결정된다.

크리틱의 학습에는 분위수 허버 손실을 사용한다. 현재 크리틱이 출력한 분위수와 TQC 절단을 거친 타깃 분위수 사이의 모든 쌍대 오차를 구성하고, 허버 손실로 매끄럽게 완화하며, 각 분위수 위치를 나타내는 τ 가중을 적용해 최종 손실을 산출한다. 허버의 전이 구간을 정하는 k 값은 L2 와 L1 사이의 민감도를 조절하여 외란과 이상치에 대한 견고성을 높인다. 액터의 목적함수는 크리틱이 추정한 기대 Q(크리틱과 분위수 평균)를 높이면서 엔트로피 항을 통해 탐험을 유지하도록 구성되며, 온도 계수 α 는 목표 엔트로피를 기준으로 자동 조절되어 상태·행동 차원과 환경 난이도에 맞춘 탐험 강도를 스스로 정렬한다.

타깃 안정화를 위해 크리틱의 타깃 네트워크를 폴리악 평균으로 갱신하여 급격한 파라미터 변화를 완화하고, 분포형 타깃의 노이즈를 줄인다. 경험 재현은 우선순위 리플레이를 사용해 TD 오차가 큰 전이를 더 자주 샘플링함으로써 표본 효율을 높인다. 한 번의 학습 스텝에서 모델은 배치를 샘플링한 후, 자동 온도 조절이 활성화된 경우 α 를 먼저 업데이트하고, 다음 상태의 정책 행동으로부터 타깃 분위수를 구축한다. 이어서 현재 크리틱의 분위수와 타깃 분위수로 분위수 허버 손실을 최소화하고, 같은 배치에서 정책 행동을 다시 샘플링해 기대 Q 와 로그확률로 액터 손실을 구성하여 정책을 갱신한다. 마지막으로 설정된 주기에 따라 폴리악 평균으로 타깃 크리틱을 갱신한다.

이와 같은 설계는 분포형 가치추정과 상위 분위수 절단이 과대추정을 제어하고, 자동 온도 조절이 탐험-이용 균형을 상황에 맞게 유지하며, 폴리악 갱신과 우선순위 리플레이가 수렴 안정성과 표본 효율을 뒷받침한다. 결과적으로 모델은 기대 성능과 위험 관리 모두에 민감하게 반응하는 정책으로 수렴한다.

실험은 약 700,000 스텝으로 진행되었다. 실험환경은 Ubuntu 22.04, ROS2 humble 에서 진행되었으며, 시뮬레이션 환경으로 gazebo 를 사용하였다. 맵의 크기는 12*12 이며, 한 에피소드가 끝이 나면 로봇의 생성 위치, 목표점, 장애물의 위치가 랜덤으로 정해지면서 환경이 리셋된다.

학습 결과, Q 값은 초반의 불안정 구간을 지나 약 1.85 만 스텝 이후 양수로 복귀해 점진적으로 상승·안정화되었고, 엔트로피 계수 α 는 초기 ≈0.712 에서 약 1.18 만 스텝 무렵 ≤0.05 로 빠르게 하락한 뒤 ≈0.018 수준에서 미세 조정되며 정책의 결정성이 강화되었다.

모델 품질을 직접 가늠하는 지표들도 일관된 개선을 보였다. 평균 Q 는 중·후반 구간에서 약 13.5- 14.3 범위로 안정화되었고, 분포 상한을 반영하는 Q_max 는 32.95 만 스텝 부근에서 28.11 로 정점을 기록해 고보상 시나리오에 대한 표현력이 확보되었음을 시사한다. Critic 손실은 대체로 0.10- 0.16 사이에서 완만히 수렴했고, Actor 손실은 −12~−13 대의 음수 구간에서 일관되게 유지되어 정책이 더 큰 Q 를 유도하는 방향으로 최적화되고 있음을 보여준다. 종합하면, 본 학습은 빠른 α 하향을 동반해 탐험에서 활용으로 매끄럽게 전환하며, 평균 가치의 안정화와 분포 꼬리의 확장을 함께 달성하는 수렴 패턴을 보였다.

또한 경로 생성 테스트 결과, 100 개의 무작위 에피소드에서 경로 생성 성공률을 92 퍼센트를 달성하였다.

Ⅲ. 결론

본 연구는 360° LaserScan 을 이용한 TQC 기반 지역 경로 계획이 다양한 환경에서 안정적으로 수렴하며 정책의 결정성과 견고성을 동시에 향상시킴을 보였다. 약 70 만 스텝 학습 동안 평균 Q 의 안정화, Q_max 의 확장, Critic 손실의 완만 수렴, 그리고 α 의 급격한 하향이 관찰되어 과대추정 억제와 탐험- 이용 균형의 자동 조절이 효과적임을 확인했다. 우선순위 리플레이·폴리악 타깃·체크포인트 운영을 결합한 실험 전략은 표본 효율과 정책 안전성을 높여 실제 로봇 적용 가능성을 뒷받침했다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 과학기술 사업화진흥원의 지원을 받아 수행된 연구임(RS-2025-25444562).

This work was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) grant funded by the Korea government(Ministry of Science and ICT) (RS-2025-25444562).

본 연구는 국방기술진흥연구소에서 지원하는 국방 중소기업 역량강화 사업(No.DC2023CS)의 연구수행으로 인한 결과물임.

This study is the result of the reserch performance of "Defense SMEs Competency Enhancement Program" (NO.DC2023CS) project supported by "Korea Research Institute for defense Technology planning and advancement"

참 고 문 헌

[1] Kuznetsov, A., Shvechikov, P., Grishin, A., & Vetrov, D. (2020, November). Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In International conference on machine learning (pp. 5556-5566). PMLR..

[2] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018, July). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning (pp. 1861-1870). Pmlr.

[3] Dabney, W., Ostrovski, G., Silver, D., & Munos, R. (2018, July). Implicit quantile networks for distributional reinforcement learning. In International conference on machine learning (pp. 1096-1105). PMLR.