TD3 기반 우선적 경험 재생 방식을 통한 병목 구간 통과 자율주행 정책 연구

엄찬인, 이동수, 권민혜*
숭실대학교

{eci0623, movementwater}@soongsil.ac.kr, *minhae@ssu.ac.kr

Autonomous Driving Strategy for Bottleneck Traffic Trough Prioritized Experience Replay with TD3

Chanin Eom, Dongsu Lee, Minhae Kwon*
Soongsil University

요 약

자율주행 기술은 인공지능 기술의 발전과 함께 큰 도약을 이루었지만, 복잡한 도로 환경에서의 주행 정책 연구는 그 중요도에 비해 비교적 적게 수행되고 있다. 본 연구에서는 복잡도가 높은 병목 도로를 성공적으로 통과할 수 있는 자율주행 정책 학습을 위한 부분 관측 가능한 마르코프 의사결정 과정(Partially Observable Markov Decision Process; POMDP)을 제안한다. 자율주행차량의 학습에는 심층강화학습 알고리즘인 Twin Delayed Deep Deterministic Policy Gradient(TD3)를 사용하며, 학습 과정에서의 경험 재생(Experience Replay) 방식으로는 우선적(Prioritized) 경험 재생을 사용한다. 결과적으로 우선적 경험 재생 방식을 통해 학습된 개체는 무작위(Random) 경험 재생 기반의 차량보다 선제적인 차선 변경을 통해 높은 속력을 유지하는 것을 확인하였다.

Ⅰ. 서 론

인공지능 기술의 발전과 함께 다양한 도로 환경에서의 자율주행 연구가 활발히 진행되고 있다. 이러한 발전에 따라 완전 자율주행 기술에 대한 관심이 높아지고 있지만, 이에 필수적인 복잡한 도로 환경에서의 자율주행 연구는 비교적 적게 수행되고 있다. 본 연구에서는 도로의 복잡도가 높은 병목 구간을 성공적으로 통과할 수 있는 POMDP를 제안하며, 정책 학습에는 TD3 알고리즘을 사용한다. 이때 학습 과정에서 중요도가 높은 경로(trajectory) 정보를 우선적으로 학습에 활용하는 prioritized 경험 재생 방식을 사용한다.

Ⅱ. 우선적 경험 재생 방식 기반 심층강화학습 모델

본 연구에서는 강화학습 문제 정의를 위해 부분적인 관측을 허용하는 POMDP 를 제안한다. POMDP 는 튜플 $\langle S, A, O, R, \gamma \rangle$로 표현되며 각각의 요소는 상태 $s_t \in S$, 행동 $a_t \in A$, 관측 $o_t \in O$ 과 보상함수 $R_t(s_t, a_t, s_{t+1})$ 및 감가율(discount factor) γ 을 의미한다. 또한, 본 연구에서는 정책 학습을 위해 경험 재사용이 가능한 off-policy 기반의 알고리즘 TD3 를 사용한다. 즉, 개체는 매 시점 t에서의 경로 정보 $\langle o_t, a_t, r_t, o_{t+1} \rangle$를 버퍼에 저장한 뒤 재사용하며, 이때 prioritized 경험 재생 방식을 사용한다.

Ⅱ.1 병목 도로 통과를 위한 Partially Observable Markov Decision Process

본 연구에서 고려하는 도로 환경은 Y개의 병합지점(merge point) $M = \{m_1, \cdots, m_Y\}$ 과 N 대의 차량 $C = \{c_1, \cdots, c_N\}$가 포함된 타원형 도로이다. 차량 집합 C는 1대의 자율주행차량 c_N과 $N-1$ 대의 일반차량 $c_i (i \neq N)$으로 구성된다.

Observation : 관측 정보 $o_t \in O$는 자율주행차량 기준 전/후방 관측 가능 거리 W 및 차선 H 내의 상태 정보로 다음과 같다.

$$o_t = \left[\Delta \boldsymbol{v}_t^T, \Delta \boldsymbol{p}_t^T, \boldsymbol{\rho}_t^T, n_{t,W}, v_{t,N}, p_{t,N}, k_{t,N}, n_{t,N} \right]^T$$

여기서 $\Delta \boldsymbol{v_t} = \left[\Delta v_{t,l_1}, \cdots \Delta v_{t,l_H}, \Delta v_{t,f_1}, \cdots, \Delta v_{t,f_H} \right]^T$ 는 차선별 전/후방 차량과 자율주행 차량 사이의 상대 속도를 의미하며,

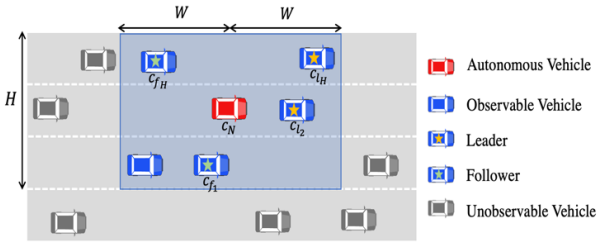

그림 1. 관측 가능 차량 정의

$\Delta \boldsymbol{p_t} = \left[\Delta p_{t,l_1}, \cdots, \Delta p_{t,l_H}, \Delta p_{t,f_1}, \cdots, \Delta p_{t,f_H} \right]^T$ 는 상대 거리를 의미한다. 이때 임의의 관측 가능 차선 $h \in [1, \cdots, H]$ 에서의 leader 차량 c_{l_h}은 차선별 관측된 전방차량 집합 C_{obs,l_h} 중 자율주행차량과의 상대거리 절댓값이 최소인 차량을 의미한다. Follower 차량 c_{f_h}은 차선별 관측된 후방차량 집합 C_{obs,f_h} 중 자율주행차량과의 상대 거리 절댓값이 최소인 차량을 의미한다(그림 1). $\boldsymbol{\rho_t^T}$는 전방 차선별 차량 밀도로 관측 가능 거리 W 대비 관측된 차량이 차지하고 있는 도로 비율을 의미한다. 즉, h 번째 차선의 차량 밀도 $\rho_{t,h}$ 는 해당 차선에 관측된 차량 대수 $|C_{obs,l_h}|$에 대해서 $\rho_{t,h} \propto \frac{|C_{obs,l_h}|}{W}$를 만족한다.

Action: 본 연구에서 개체의 행동 $a_t = \{a_{t,acc}, a_{t,lc}\}$은 가속도 조절 $a_{t,acc}$ 및 차선변경 $a_{t,lc}$ 행동으로 구성된다. 이때 가속도 조절 $a_{t,acc} \in [a_{min}, a_{max}]$은 최소 가속도 a_{min}와 최대 가속도 a_{max} 의 범위 내에서 연속적인 값을 선택한다. 차선 변경 행동 $a_{t,lc} \in \{-1, 0, 1\}$ 은 이산적인 행동 공간에서 정의되며 -1 은 오른쪽, 1은 왼쪽으로의 차선 변경을 의미한다. 개체는 $a_{t,lc} = 0$일 경우에는 차선을 유지한다.

Reward: 보상 r_t는 현재 상태 s_t, 현재 행동 a_t 그리고 다음 상태 s_{t+1}에 대한 함수 $R_t(s_t, a_t, s_{t+1})$형태이며, 다음과 같다.

$$R_t(s_t, a_t, s_{t+1}) = \eta_1 \mathcal{R}_1 + \eta_2 \mathcal{R}_2 + \eta_3 \mathcal{R}_3 + \eta_4 \mathcal{R}_4 \quad (1)$$

$\mathcal{R}_1$는 목표속도 준수를 위한 항으로 목표속도 v^* 및 제한 속도 v_{limit}에 대해 다음과 같이 정의한다.

* 교신저자(Corresponding Author)

표1. 경험 재생 방식 별 차선 변경 경향성 분석

	$\Delta p_{t,l}$	$v_{t+1,N}$
Prioritized	20.38 m	8.20 m/s
Random	24.08 m	5.32 m/s

$$\mathcal{R}_1 = \begin{cases} \frac{v_{t+1,N}}{v^*} & v_{t+1,N} \le v^* \\ \frac{v_{limit}-v_{t+1,N}}{v_{limit}-v^*} & v_{t+1,N} > v^* \end{cases} \quad (2)$$

즉, 개체는 v^*에 근접한 주행을 할 시 최대의 보상을 획득하며, v_{limit}을 초과한 속력을 낼 시 음의 보상을 받는다.

두 번째 항 $\mathcal{R}_2 = |a_{t,lc}|(\Delta p_{t+1,l} - \Delta p_{t,l} - \delta_{lc})$ 는 성공적인 차선 변경을 위한 보상항으로 개체가 차선 변경을 수행한 경우($|a_{t,lc}| = 1$)에 적용된다. 여기서, $\Delta p_{t,l}$와 $\Delta p_{t+1,l}$은 개체의 차선 변경 수행 전, 후 동일 차선 leader와의 상대 거리를 의미하며, $\delta_{lc} \in (0, W]$는 성공적인 차선 변경 기준을 결정하는 임곗값(threshold)을 의미한다. 구체적으로, 개체가 δ_{lc}이상의 상대 거리 이득($p_{t+1,l} - \Delta p_{t,l} > \delta_{lc}$)을 얻는 차선 변경을 수행했을 경우를 성공적인 차선 변경으로 해석하여 양의 보상을 획득한다.

이어서 $\mathcal{R}_3 = |a_{t,lc}| \times \min\left[0,\ 1 - \left(\frac{s^*}{\Delta p_{t+1,f}}\right)^2\right]$ 은 동일 차선 follower의 안전거리 s^*를 침범하는 행동을 약화한다. 여기서, $\Delta p_{t+1,f}$ 는 동일 차선 follower와의 상대 거리를 의미한다. 안전거리 s^*는 Intelligence Driving Module (IDM)[3] 컨트롤러에 기반하여 계산한다. 즉, $p_{t+1,f} < s^*$ 인 경우를 follower의 안전거리를 침범한 행동으로 간주하여 페널티를 부여한다.

$\mathcal{R}_4$ 는 개체가 수행 불가한 차선 변경 행동을 결정했을 때 부여하는 페널티를 의미한다.

Ⅱ.2 우선적 경험 재생 방식을 통한 정책학습

Prioritized 경험 재생 방식은 버퍼에 저장된 경로 정보의 우선순위 p_d 에 기반하여 해당 정보의 샘플링 확률 $P(d)$ 을 결정하는 방식으로 샘플링 확률 $P(d)$는 다음과 같이 정의한다.

$$P(d) = \frac{p_d^{\omega_{pr}}}{\sum_{d=1}^{\mathcal{D}} p_d^{\omega_{pr}}} \quad (3)$$

이때 $\mathcal{D}$는 버퍼 크기를 의미하며, ω_{pr}는 우선순위 규칙을 적용할 정도를 의미한다. 우선순위 p_d는 해당 정보의 TD-error인 δ_d를 통해 결정되며 다음과 같이 정의한다.

$$p_d = |\delta_d| + \epsilon = |y_d - Q_{\theta_i}(o_d, a_d)| + \epsilon \quad (4)$$

여기서 ϵ는 $P(d)$의 분모가 0 이 되지 않게 하기 위한 충분히 작은 상수이다. 또한, $y_d = r_d + \gamma \min_{i=1,2} Q_{\theta_i'}(o_{d+1}, \pi_{\phi'}(o_{d+1}) + \epsilon_{TD3})$ 는 정책 평활화(smoothing)를 위한 잡음 ϵ_{TD3}이 적용된 TD3 의 크리틱 TD-target 을 의미한다. Q_θ와 π_ϕ는 Q 값 및 정책 근사를 위한 크리틱 네트워크와 액터 네트워크를 의미한다.

Prioritized 경험 재생 방식에서는 우선순위가 갱신될 때마다 변경되는 샘플링 확률분포를 정정하기 위해 Importance Sampling (IS) 가중치 $w_d = \left(\frac{1}{\mathcal{D}} \cdot \frac{1}{P(d)}\right)^{\omega_{is}}$ 를 고려한다. 이때 ω_{is} 는 IS 가중치를 고려할 정도를 의미한다. IS 가중치는 TD3 크리틱 네트워크의 목적함수 $J(\theta_i)$ 에서 고려되며 이는 B 개만큼 샘플링된 배치 내 경로 정보 j에 대해서 다음과 같이 정의한다.

$$J(\theta_i) = \frac{1}{B}\sum_{j=1}^{B} w_j \left(y_j - Q_{\theta_i}(o_j, a_j)\right)^2 \quad (5)$$

Ⅲ. 모의실험 설정 및 경험 재생 방식에 따른 주행 성능 비교

본 연구에서는 개체 학습 및 성능 평가를 위해 교통제어 시뮬레이터인 FLOW[4]를 사용한다. 이때, 도로 내 전체 차량 대수 N은 32 대로 설정하였으며, 자율주행차량은 1 대로 설정하였다. 도로 내 병합지점인 Y는 2 개로 차선이 감소하는 구간과 증가하는 구간을 포함하는 타원형 도로를 고려한다.

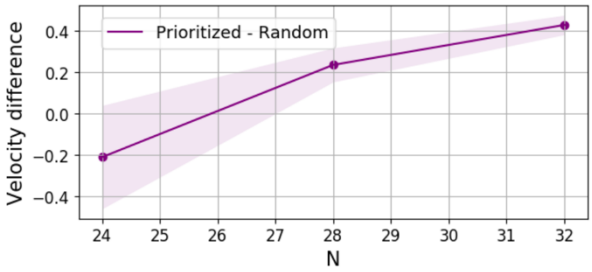

그림 2. 경험 재생 방식 별 도로 구간 별 평균 속력차이 비교

표 1은 경험 재생 방식에 따른 차선 변경 경향성을 측정한 결과이다. 표에서 $\Delta p_{t,l}$ 은 차선 변경 전 동일 차선 leader와의 상대 거리를 의미하고, $v_{t+1,N}$ 는 차선 변경 후 개체의 속력을 의미한다. 해당 표를 통해 prioritized 방식의 개체가 random 방식의 개체보다 선제적인 차선 변경을 수행함을 확인할 수 있다. 선제적인 차선 변경을 수행하는 차량은 병목 도로 내에서 동일 차선 leader의 속력 감소에 영향을 적게 받으므로, 비교적 높은 속력을 유지할 수 있다. 따라서 표 1의 $v_{t+1,N}$ 결과 또한, prioritized에서 높은 것을 확인할 수 있다.

그림 2는 도로 내 전체 차량 대수 N 이 증가함에 따른 prioritized 및 random 기반 개체의 전 구간 평균 속력 차이를 나타낸 그래프이다. 그래프에서 실선은 경험 재생 방식 간 속력 차이의 평균값이며, 음영은 표준편차를 의미한다. 그래프를 통해 도로가 복잡해질수록 prioritized 방식의 개체가 더 빠른 주행을 수행함을 확인할 수 있다. 이는 prioritized 기반의 자율주행차량이 복잡한 도로에서 우수한 주행을 수행함을 의미한다.

Ⅳ. 결론

본 연구에서는 차량 정체가 빈번히 발생하는 병목구간을 성공적으로 통과하기 위한 POMDP 를 제안하였다. 또한 TD3 를 통한 학습 과정에서 prioritized 경험 재생 방식을 적용하였다. 결과적으로, 경험 재생 방식의 차이는 복잡한 도로를 통과하기 위한 정책 학습에 유의미한 차이를 도출하는 것을 확인하였다. prioritized 방식의 개체는 random 방식의 개체보다 선제적인 차선 변경을 수행함으로써 비교적 높은 속력을 유지할 수 있는 것을 확인하였다. 또한 prioritized 기반의 개체가 도로의 복잡도 증가에도 강건한 주행이 가능함을 확인하였다.

사 사

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단 (NRF-2020R1F1A1069182) 및 정보통신기획평가원(2021-0-00739, 분산/협력 AI 기반 5G+ 네트워크 데이터 분석 기능 및 제어 기술 개발)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] S. Fugimoto, H. van Hoof, et al., “Addressing function approximation error in actor-critic methods,” ICML, 2018.

[2] T. Schaul, J. Antonoglou, et al., “Prioritized experience replay,” ICLR, 2016.

[3] M. Treiber, A. Hennecke, et al., “Congested traffic states in empirical observations and microscopic simulation,” Physical review E, vol. 62, no. 2, pp. 1805-1824, 2000.

[4] C. Wu, A. R. Kreidieh, et al., “Flow: A modular. learning framework for mixed autonomy traffic,” IEEE Transactions on Robotics, vol. 38, no. 2, pp. 1270-1286, 2021.