심층강화학습 기반 자율주행차량의 차선 변경 정책 안정성 평가

이동수, 권민혜*
숭실대학교
movementwater@soongsil.ac.kr, *minhae@ssu.ac.kr

Stability Evaluation on Lane-changing Policies for Deep Reinforcement Learning-based Autonomous Vehicles

Dongsu Lee, Minhae Kwon*
Soongsil University

요 약

자율주행기술과 같은 고신뢰 서비스(mission-critical service)는 상용화를 위해서는 학습 결과물의 안정성이 매우 중요하다. 본 연구에서는 자율주행차량의 안정적인 차선변경 정책을 위한 마르코프 의사결정(Markov Decision Process; MDP) 모델을 제안하고 대표적인 심층강화학습 알고리즘인 TD3와 PPO를 이용하여 자율주행차량을 학습시킨다. 학습된 자율주행차량의 속도분포의 엔트로피(entropy)를 확인하여 각 알고리즘별 학습 결과물의 안정성을 평가하였다. 그 결과 TD3를 기반으로 학습된 차량은 PPO 기반 학습차량에 비해 엔트로피(entropy) 값이 약 0.4 배 수준으로 낮아 더 안정적인 속도제어가 가능함을 확인하였다.

Ⅰ. 서 론

최근 인공지능 기술의 발전과 함께 자율주행기술의 연구가 활발히 진행되고 있다. 자율주행기술과 같은 고신뢰 서비스의 상용화를 위해서는 주행 효율성 뿐만 아니라 안정성 역시 중요한 척도가 된다. 본 연구에서는 차선 변경 학습을 위한 MDP 모델을 대표적인 심층강화학습 기반의 알고리즘인 Twin Delayed DDPG (TD3)[1] 와 Proximal Policy Optimization (PPO)[2]을 통해 자율주행차량을 학습시키고자 한다[3]. 그리고 불안정성 지표인 엔트로피를 이용하여 학습된 차량의 주행 정책 안정성을 비교한다.

Ⅱ. 차선 변경 학습을 위한 심층강화학습 문제 정의

모든 강화학습의 기본 동작은 마르코프 의사결정 과정(Markov Decision Process; MDP)으로 정의할 수 있다. MDP는 환경에 대한 완벽한 상태 정보를 획득할 수 있다는 가정이 존재하는데, 실제 환경에서는 부분적이거나 잡음(noise)이 섞인 불완벽한 상태 정보를 얻게 된다. 이러한 문제를 해결하기 위해 제안된 모델이 Partially Observable MDP (POMDP)이다. POMDP는 $< S, A, O, R, \gamma >$의 튜플로 표현할 수 있다. 튜플은 시간 t에서 유한한 상태 공간 집합 $s_t \in S$, 유한한 행동 공간 집합 $a_t \in A$, 유한한 관측 공간 집합 $o_t \in O$, 행동에 따른 보상 $R(s_t, a_t, s_{t+1})$, 마지막으로 시간에 따른 감가율 γ을 포함한다.

Ⅱ.1. 도로 환경의 정의

본 논문에서는 차선변경이 필요한 2차선 원형 도로에서 한 대의 자율주행 차량과 여러 대의 비 자율주행차량이 혼재된 도로상황을 고려한다. 도로의 길이는 l이며 도로 내 차량의 집합 E는 자율주행차량 $[e_N]$과 비 자율주행차량 $[e_1, e_2 \cdots, e_{N-1}]$을 포함한다. 이와 같은 환경의 시간 t에서의 상태 정보는 $3N$차원으로 다음과 같이 정의 된다($s_t \in \mathbb{R}^{3N \times 1}$).

$$s_t = [v_1, p_1, k_1, v_2, p_2, k_2, \cdots, v_N, p_N, k_N]$$

여기서 v_n는 n 번째 차량의 속도, p_n은 n번째 차량의 위치, 마지막으로 k_n은 n 번째 차량이 위치한 차선 번호를 의미한다. 만약 특정 차량 e_n이 원형도로의 가장 바깥쪽 차선에 있다면 $k_n = 0$으로 정의하고 안쪽 차선으로 이동함에 따라 차선 번호가 1씩 증가한다.

Ⅱ.2. POMDP 정의

개체는 학습을 위해 시간 t에서의 환경 상태 정보 s_t를 관측하여 부분 관측 정보 o_t를 획득한다. 본 연구에서 자율주행차량 e_N이 관측 가능한 차량의 집합 $E_{obs} = [e_{l_0}, e_{l_1}, e_{f_0}, e_{f_1}]$전방 차량(leader) 두 대 e_{l_0}, e_{l_1}와 후방 차량(follower) 두 대 e_{f_0}, e_{f_1} 총 네 대의 차량을 포함한다. 이때 관측 정보 o_t는 10차원으로 다음과 같이 정의 된다($o_t \in \mathbb{R}^{10 \times 1}$).

$$o_t = [v_{f0}, v_{f1}, v_{l0}, v_{l1}, v_N, P_{f0}, P_{f1}, P_{l0}, P_{l1}, k_N]$$

여기서 P는 자율주행차량의 위치 p_N과 관측 가능한 차량의 위치 p_{obs}의 상대거리를 의미한다.

개체가 매 시간 t마다 수행하는 행동 $a_t = \{a_{acc}, a_{lc}\}$는 가속도 조절 a_{acc}과 차선 변경 a_{lc}을 포함한다. 이때 a_{acc}는 유한하며 연속적인 실수 공간 $a_{acc} \in [acc_{\min}, acc_{\max}]$에서 정의 된다. $acc_{\min}$은 자율주행차량이 수행할 수 있는 최소 가속도를, $acc_{\max}$는 최대 가속도를 의미한다. a_{lc}는 유한하며 이산적인 실수 공간 $a_{lc} \in \{-1, 0, 1\}$에서 정의 된다. $a_{lc} = 0$인 경우 자율주행차량은 차선을 유지하며, $a_{lc} = -1$은 바깥쪽(우측)으로의 차선 변경을, 그리고 $a_{lc} = 1$는 안쪽(좌측)으로의 차선 변경을 의미한다.

개체는 주어진 상태를 관측한 뒤 행동을 선택하고 이를 바탕으로 보상 R_t을 얻게 된다. 차선 변경 및 가속도 조절 학습을 위한 보상함수는 다음과 같다.

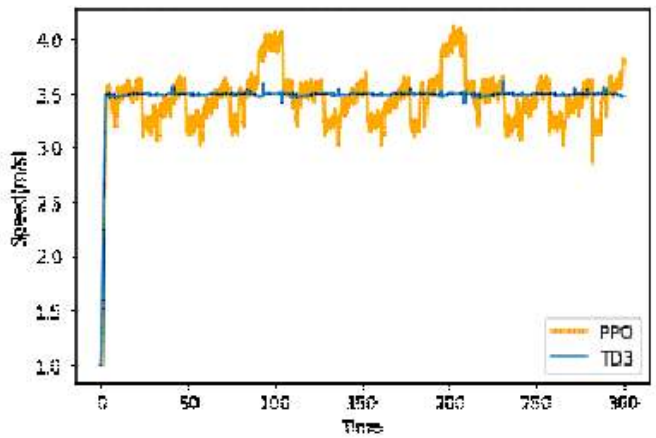

그림 1. 시간에 따른 자율주행차량의 속도 변화

$$R_t = \eta_1\left(1-\left|\frac{v_{t+1,N}-v^*}{v^*}\right|\right)+\eta_2\left(\min\left[0,1-\left(\frac{s^*}{P_{t+1,f}}\right)\right]\right) + \eta_3 \times |a_{lc}| \min(0, P_{t+1,l}-P_{t,l}) \times P_{t,l} \quad (1)$$

보상함수는 세 가지 항으로 구분할 수 있다. $\left(1-\left|\frac{v_{t+1,N}-v^*}{v^*}\right|\right)$는 가속도 조절 후 자율주행차량의 속도$v_{t+1,N}$와 목표속도 v^*를 이용하여 원하는 속도로 주행할 수 있도록 한다. $\left(\min\left[0,1-\left(\frac{s^*}{P_{t+1,f}}\right)\right]\right)$는 차선 변경을 수행한 후 차량의 후방차량과의 상대거리 $P_{t+1,f}$와 안전거리 s^*를 고려한 안전거리 계수이다[4]. $|a_{lc}| \min(0, P_{t+1,l}-P_{t,l}) \times P_{t,l}$는 차선 변경을 수행한 경우 t시간에서 선두차량과의 상대거리 $P_{t,l}$와 $t+1$시간에서 선두차량과의 상대거리 $P_{t+1,l}$비교하여 의미 없이 수행하는 차선 변경을 제어한다.

본 논문에서는 제시한 MDP 문제 해결을 위해 심층강화학습 알고리즘인 TD3와 PPO를 사용한다. 개체는 특정 상태에서 최적의 행동을 결정하는 정책을 학습한다. 각각의 정책 평가 및 업데이트에는 관측-행동 가치 함수 $Q(o,a)$및 관측 가치 함수 $V(o)$가 이용된다.

Ⅲ. 성능평가

본 연구에서는 앞서 제안한 POMDP문제를 풀기 위해 대표적인 심층강화학습 알고리즘인 TD3 및 PPO를 기반으로 자율주행차량을 학습 시킨다. 그 후, 알고리즘 별 학습 차량의 주행 정책의 안정성을 비교평가 하고자 한다.

Ⅲ.1 모의실험 설정

자율주행차량의 성능 평가를 위해 심층 강화학습 라이브러리와 교통 시뮬레이터를 통합한 프레임워크 FLOW[5]를 이용한다. 도로는 2차선 원형도로이며 $l=260m$이다. 전체 차량 $|E|$대비 자율주행차량 $|E_{non}|=8$이다. 이때 비 자율주행차량은 도로 내에서 일정한 간격으로 배치하였으며 Intelligence Driving Model (IDM) 컨트롤러[6]를 이용하여 주행한다. 모든 비 자율주행차량은 $1m/s$의 등속주행을 한다. 자율주행차량의 목표속도 $v^*=3.5m/s$,안전거리 s^*는 IDM 컨트롤러에서 제공하는 함수로 정의하며, 가속도 공간은 $acc_{\min}=-1m/s^2, acc_{\max}=1m/s^2$로 정의한다.

Ⅲ.2 자율주행차량의 주행 정책 안정성 평가

성능 평가에는 평균 속도 및 속도 분포에 대한 엔트로피를 이용하여 전체 에피소드에서 자율주행차량 정책의 안정성을 정량화하여 평가하였다. 엔트로피 H는 shannon entropy[7]를 이용하였다.

$$H=\sum_i (p_i)\ln\left(\frac{1}{p_i}\right)$$

여기서 p는 개체가 경험한 주행 속도에 대한 확률을 의미한다.

표 1. 5개의 랜덤시드로 학습시킨 자율주행차량의 속도 및 엔트로피에 대한 평균과 분산

Algorithm	Speed (m/s)		Entropy (H)	
	mean	variance	mean	variance
PPO	$3.94m/s$	$0.9\times10^{-}$	4.87	0.13
TD3	$3.48m/s$	$0.4\times10^{-}$	2.05	0.13

주행 정책 안정성 평가는 그림 1의 학습 차량의 시간 변화에 따른 속도 변화 및 표1에서 5개의 다른 랜덤시드로 학습시킨 차량의 평균속도, 평균 엔트로피, 엔트로피의 분산을 통해 확인할 수 있다. TD3의 경우 에피소드 전체에서 큰 속도의 변화 없이 목표 속도인 $3.5m/s$를 유지한 채 주행하는 모습을 확인하였다. 반면 PPO의 경우 TD3와 비교하였을 때 속도의 변화 폭이 큰 것을 확인할 수 있다. 또한 PPO로 학습시킨 차량의 H는 TD3에 비해 2.4배 정도의 불안정성 척도를 보인다(표 1). 이는 PPO로 학습시킨 자율주행차량의 경우 일관된 정책이 아닌 불안정한 정책을 통해 주행한다는 것을 의미한다. 반면 TD3로 학습한 자율주행차량은 설정된 목표속도 $3.50m/s$에서 큰 변동이 없으며 낮은 엔트로피 값으로 안정적인 정책으로 주행하는 모습을 확인하였다.

Ⅳ. 결론

본 논문에서는 차선 변경을 위한 POMDP 모델을 제안하였다. 제안된 POMDP 모델로 디자인된 문제는 심층강화학습 알고리즘인 TD3와 PPO를 이용하여 성공적으로 해결하였다. 또한, 두 심층강화학습 알고리즘의 성능 차이를 비교하였다. TD3로 학습시킨 차량의 경우 PPO로 학습시킨 차량에 비해 목표속도 유지 능력이 뛰어남을 확인할 수 있었다. 결과적으로 논문에서 제안한 MDP 모델에 대해 TD3는 PPO보다 엔트로피가 0.4 배 낮은 수준으로 보다 안정적인 주행 정책을 수립함을 확인하였다.

사 사

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업(IITP- 2021-2020-0-01602)과 한국 연구재단(NRF- 2020R1F1A1069182)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," ICML, 2018.

[2] J. Schulman, F. Wolski, et al., "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347, 2017.

[3] 이동수, 권민혜, "심층강화학습기반 자율주행차량을 이용한 원형도로의 stop-and-go wave 현상 해결 전략 연구," 한국통신학회 논문지, vol.46, no.10, 2021.

[4] 이동수, 권민혜, "PPO기반 자율주행차량의 효율적이고 안전한 차선 변경 정책 연구," 한국통신학회 하계종합학술대회, 2021.

[5] C. Wu, A. Kreidieh, et al., "Flow: Architecture and Benchmarking for Reinforcement Learning inTraffic Control," arXiv preprint arXiv:1710.05465, 2017.

[6] M. Treiber, A. Hennecke, et al., "Congested Traffic States in Empirical Observations and Microscopic Simulations," Physical Review E, vol.62, no.2, pp. 1805-1824, 2000.

[7] C. E. Shannon, "A Mathematical Theory of Communication," The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, 1948.