https://doi.org/10.7840/kics.2023.48.6.690

우선적 경험 재생 방식을 이용한 병목 구간 통과 자율주행 정책 연구

엄 찬 인•, 이 동 수*, 권 민 혜°

Autonomous Driving Strategy for Bottleneck Traffic with Prioritized Experience Replay

Chanin Eom•, Dongsu Lee*, Minhae Kwon°

요 약

인공지능을 활용한 자율주행 연구가 가속화됨에 따라, 도로 정체와 같은 복잡한 환경에서 주행 가능한 자율주행 기술에 대한 수요가 증가하고 있다. 이에 고차원의 상태정보에 즉각적인 의사결정이 가능한 심층강화학습(deep reinforcement learning) 기반의 자율주행 연구가 주목을 받고 있다. 본 연구에서는 교통 정체가 빈번히 발생하는 병목구간의 성공적인 통과를 위한 부분 관측가능한 마르코프 의사결정과정(Partially Observable Markov Decision Process; POMDP)을 제안한다. 정책 학습에는 Twin Delayed Deep Deterministic Policy Gradient(TD3) 알고리즘을 사용하며, 우선적 경험 재생(prioritized experience replay) 기반의 샘플링 방식을 사용한다. 결과적으로 우선적 경험 재생 기반의 자율주행차량이 무작위(random) 경험 재생 기반 개체보다 복잡한 도로에서 우수한 성능을 보임을 확인하였다.

Key Words : Autonomous driving system, Bottleneck traffic, Deep reinforcement learning, Partially observable Markov decision process, Twin delayed deep deterministic policy gradient, Prioritized experience replay

ABSTRACT

Recently, the demand for a higher level of autonomous driving technology in complex circumstances has increased. Deep reinforcement learning is gaining attention as a promising solution that enables instant decision-making based on high-dimensional state information. In this study, we propose a Partially Observable Markov Decision Process (POMDP) to train the autonomous driving policy that can successfully navigate bottleneck congestion. Furthermore, we suggest using the prioritized experience replay method in Twin Delayed Deep Deterministic Policy Gradient (TD3) to train the policy. As a result, we confirm that the vehicles trained with the prioritized experience replay method outperform the vehicles trained with the random experience replay method.

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

• First Author : Soongsil University Department of Information and Telecommunication Engineering, eci0623@soongsil.ac.kr, 학생회원

° Corresponding Author : Soongsil University School of Electronic Engineering, minhae@ssu.ac.kr, 종신회원

* Soongsil University Department of Information and Telecommunication Engineering, movementwater@soongsil.ac.kr, 학생회원

논문번호 : 202302-027-C-RN, Received February 13, 2023; Revised March 21, 2023; Accepted April 6, 2023

Ⅰ. 서 론

최근 자율주행 기술은 인공지능 연구의 발전과 함께 빠르게 발전하고 있다. 실제로 다양한 도로 시나리오에서 인공지능을 적용한 주행 정책 학습 연구가 많이 진행되고 있으며 이에 따라 완전한 자율주행 기술에 대한 관심이 높아지고 있다. 하지만, 완전 자율주행으로 나아가기 위해 필수적인 불규칙한 도로 환경에서의 자율주행 연구는 비교적 적게 수행되고 있다. 이러한 경향성은 산업 분야에서도 확인할 수 있는데, 현재 국내에서 정체 현상 고려가 가능한 3단계[1] 이상의 자율주행 기술은 전체 대비 매우 작은 비율로 추정된다[2]. 이는 복잡한 도로를 고려하는 자율주행 기술 연구의 필요성을 입증한다.

자율 주행 기술은 차량의 센싱 모듈을 통해 주변 도로 환경을 인식하는 인지 단계, 인식된 정보를 통해 주행 경로 등을 예측하는 판단 단계, 예측된 경로를 따라 차량을 이동시키기 위한 제어 단계로 크게 3가지 단계로 구분될 수 있다. 여기서 자율주행차량의 주행 의사결정은 판단 단계에서 수행되며, 판단 단계의 기술은 크게 제어이론(control theory) 및 심층신경망(deep neural network) 기반의 기술로 분류할 수 있다.

제어이론 기반 방식이란 차량이 수학적으로 모델링된 컨트롤러(controller)를 통해 의사결정을 수행하는 방법을 의미한다. 이와 같은 방식은 차량의 행동 수행을 설명하기 용이하다는 장점이 있다. 하지만, 모든 도로 환경을 고려한 컨트롤러의 설계가 제한되기 때문에 예상치 못한 도로 환경에서의 대응이 어렵다. 또한, 자율주행차량의 차선 변경, 가속도 조절 등 독립된 행동 수행을 위해서는 개별적인 컨트롤러를 설계해야 한다는 한계가 존재한다.

반면, 심층 신경망 기반 방식은 인지 단계에서 수집된 정보를 신경망의 입력값으로 사용하여 차량 제어에 필요한 값을 도출하는 방식으로, 신경망을 통한 함수 근사(approximation)가 이루어지기 때문에 복잡한 도로 상황에서도 성공적인 주행이 가능하다. 따라서 복잡한 도로 환경을 고려하는 자율주행 시스템을 위해서는 예상치 못한 도로 환경에 대응할 수 있는 신경망 기반의 방법을 사용하는 것이 적절하다. 이에 도로 상태정보에 대해 즉각적인 의사결정이 가능한 강화학습에 신경망 기술을 결합한 심층강화학습 기반의 자율주행 연구가 주목을 받고 있다.

본 연구는 병목 구간으로 인해 정체 현상이 빈번히 발생하는 도로를 성공적으로 통과할 수 있는 자율주행 정책의 학습을 목표로 한다. 이때, 자율주행 차량은 차선 변경 및 가속도 조절의 행동을 수행하며 심층강화학습을 기반으로 자율주행 정책을 학습한다. 본 연구에서는 강화학습 문제 정의를 위한 POMDP를 제안하며 정책 학습에는 off-policy 기반 심층 강화학습 알고리즘인 TD3[3]를 사용한다. 또한 TD3 학습 과정에서의 경험 재생 방법에 따른 성능 평가를 수행한다. 구체적으로, prioritized 경험 재생[4] 방식과 random 경험 재생 방식을 통해 학습된 자율주행차량의 주행 특성 및 성능 차이를 분석한다.

본 논문의 구성은 다음과 같다. 먼저 Ⅱ장에서 본 연구와 관련된 강화학습 및 자율주행 선행 연구에 대해 살펴본 뒤, Ⅲ장에서는 병목구간 통과 정책을 위한 POMDP 제안 및 자율주행 정책 학습 방법에 대해 소개한다. 이후 Ⅳ장에서는 자율주행 정책 학습을 위한 시뮬레이션 설정 소개 및 성능평가 결과를 확인한다. 마지막으로, Ⅴ장에서는 본 연구의 결론을 맺는다.

Ⅱ. 선행연구

2.1 Model-free 기반 심층강화학습 알고리즘

심층 강화학습은 환경 내 상태전이 확률을 나타내는 모델의 사용 유무에 따라 model-based[5,6] 및 model-free로 구분할 수 있다. 불확정성이 높은 실제 세계에서는 정확한 전이 모델을 예측하기 어렵기 때문에 대부분의 강화학습 문제는 model-free 방식을 통해 해결되고 있다. Model-free 알고리즘의 종류로는 행동가치 함수(Q-function)를 근사하는 가치 기반(value-based) 방식, 개체의 정책(policy)을 직접적으로 근사하는 정책 기반 방식이 있으며, 두 가지를 결합한 액터-크리틱(actor-critic) 방식이 있다.

가치 기반 방식은 DQN(Deep Q-network)[7]을 시작으로 발전하였다. DQN에서는 심층 신경망을 활용하여 Q-value의 근사를 통한 정책 학습에 성공하였다. DQN에서 발전한 대표적인 가치 기반 알고리즘으로는 DDQN(Double DQN)[8]이 있다. DDQN은 업데이트를 적용하는 목표(target) 네트워크와 행동 선택을 위한 네트워크를 분리함으로써, 기존의 DQN에서 문제가 되었던 과대 추정(over estimation) 문제를 완화하였다.

대표적인 정책 기반 알고리즘인 REINFORCE[9]의 경우 몬테카를로(Monte Carlo; MC)[10] 방식을 통해 추정한 누적 보상 값에 대해 경사 상승법(gradient ascent)을 적용하여 정책 결정을 위한 네트워크의 직접적인 업데이트가 가능함을 보였다.

액터-크리틱 방식은 개체의 행동 결정을 위한 액터

와 행동 평가를 위한 크리틱을 통해 정책을 학습하는 방식으로 가치 함수와 정책을 동시에 고려하는 방법이다. 대표적인 알고리즘으로는 연속적인 행동 공간을 고려하는 DDPG(Deep Deterministic Policy Gradient)[11]와 TD3가 있다. 또한, 액터의 목적함수에 엔트로피를 적용함으로써 개체의 탐험(exploration)을 장려하는 SAC(Soft Actor Critic)[12]와 같이 다양한 알고리즘에 관한 연구가 진행되고 있다.

2.2 강화학습을 활용한 자율주행 연구

자율주행 기술은 시간에 따라 변화하는 복잡한 도로 상황에 대한 즉각적인 의사결정이 요구된다. 때문에 많은 자율주행 연구는 심층 강화학습 알고리즘 적용을 기반으로 진행[13] 되며, 이를 위한 MDP (Markov Decision Process) 및 POMDP를 제안하는 과정이 포함된다. [14]에서는 원형 도로에서 자율주행 차량의 가속도 조절만을 통해 stop-and-go wave 현상을 해결할 수 있음을 보였다. 차선 변경을 고려하는 자율주행 연구로는 커리큘럼 설계를 통해 차량이 성공적인 추월 정책을 학습하는 연구[15] 및 Q-learning을 통해 가속도 조절 및 차선 변경 정책을 학습하는 연구가 있다[16]. 또한, [17]에서는 안정적인 주행을 위해 차선 변경 및 가속도 조절 행동 학습을 위한 POMDP의 보상 항에서 advanced driver assistance system과 관련한 항을 고려하였다. 이를 통해 정책 수행 단계에서 추가적인 안전 모듈의 적용 없이도 높은 안정성을 지닌 주행이 가능함을 보였다.

이처럼 강화학습 기반의 자율주행 연구는 다양한 방향으로 제안되고 있다. 하지만, 대부분의 연구에서 도로 구조 및 차량 밀도 등의 복잡도는 비교적 낮게 설정되었다. 이와 같은 연구들은 자율주행차량이 정책 학습 과정에서 복잡한 도로를 경험하지 않기 때문에 차량 정체와 같이 불규칙한 환경이 조성될 경우 성공적인 의사결정을 보장하기 어렵다. 이에 본 연구에서는 개체가 복잡한 도로에서도 성공적인 주행이 가능하도록 병목 구간이 포함된 도로 환경을 고려한다.

2.3 정체 현상을 고려하는 자율주행 연구

정체 현상을 고려하는 자율주행 연구는 병목 및 병합 구간이 포함된 도로 환경을 고려한다. 하지만 대부분의 연구는 개별적인 자율주행 차량이 아닌 도로 시스템을 제어하는 방식을 해결책으로 제안하고 있다. 대표적으로 도로 구간별 교통 정보를 기반으로 특정 도로의 속도제한을 변경하는 가변속도 제한(Variable Speed Limit; VSL) 시스템 기반 연구[18,19] 및 신호등 제어 기반의 연구[20]가 있으며, 통합된 도로 환경에서 VSL과 신호등 제어를 모두 고려하는 연구[21]가 수행되고 있다. [22]는 실존하는 병목 도로를 시뮬레이터를 통해 재현하였으며, 분할된 도로의 교통정보를 바탕으로 자율주행차량의 가속도를 조절한다. 위와 같은 연구에서는 상태정보 및 행동 수행을 위한 통신 수단으로 V2X (Vehicle to Everything) 기술을 필연적으로 사용한다. 이는 통신 지연 등의 문제에 영향을 받을 수 있다는 우려가 존재한다.

이에 본 연구에서는 개체의 관측 가능 범위 내의 정보만을 활용한 POMDP를 설계한다. 즉, 상태 정보 수집 주체 및 제어 대상을 시스템이 아닌 자율주행차량으로 설정함으로써 개별 개체의 의사결정 방법에 대한 연구를 수행한다.

Ⅲ. 성공적인 병목구간 통과를 위한 심층 강화학습 모델 설계 및 정책 학습

대부분의 강화학습 문제는 MDP를 통해 모델링할 수 있다. 이때, MDP는 개체가 모든 상태정보를 관측 가능하다고 가정하기 때문에 현실적인 강화학습 모델링을 위해 부분적인 관측을 통한 의사 결정이 가능한 POMDP 모델을 사용할 수 있다. POMDP는 튜플 $< S, A, O, R, \gamma >$로 표현할 수 있으며, 여기서 $s_t \in S$는 환경의 상태(state), $a_t \in A$는 개체의 행동(action), $o_t \in O$는 특정 상태 s_t에서의 관측 가능 정보(observation)를 의미한다. $R(s_t, a_t, s_{t+1})$는 개체의 행동에 대한 보상 함수(reward function)이며, $\gamma \in [0,1)$는 시간에 따른 감가율(discount factor)을 의미한다. 본 연구에서는 현실적인 도로 환경을 고려하기 위해 부분적 관측 정보 기반의 POMDP를 제안한다.

3.1 강화학습을 활용한 자율주행 연구

본 연구에서 고려하는 도로 환경은 개의 병합지점(merge point) $M = \{m_1, \cdots, m_Y\}$과 N대의 차량 $C = \{c_1, \cdots, c_N\}$이 포함된 타원형 도로이다(그림 1). 해당 환경에서 도로 내 차량 대수 N이 증가할수록 도로의 복잡도는 증가하며, 특히 차선이 감소하는 병합 지점의 경우 병목현상으로 인한 교통정체가 발생한다. 학습 과정에서 교통정체를 경험한 개체는 혼잡지역 통과를 위해 차선 변경 등의 행동을 수행함에 따라 주행 정책을 학습할 수 있다. 차량 집합 C는 1대의 자율주행차량(Autonomous Vehicle; AV)과 $N-1$대의 일반차량(Human Vehicle; HV)으로 구성되며 차량

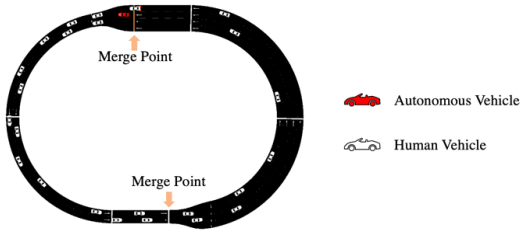


그림 1. 병목구간이 포함된 도로 구조
Fig. 1. Road structure with bottleneck traffic

종류는 도로 내 임의의 차량 c_i에 대해 다음과 같이 정의한다.

$$c_i = \begin{cases} HV & i \neq N \\ AV & i = N \end{cases}$$

위와 같은 환경에서 상태정보 s_t는 $4N$만큼의 차원을 가지며, 아래와 같이 정의한다.

$$s_t = [\boldsymbol{v_t}^T, \boldsymbol{p_t}^T, \boldsymbol{k_t}^T, \boldsymbol{n_t}^T]^T$$

여기서 $\boldsymbol{v_t} = [v_{t,1}, v_{t,2}, \cdots, v_{t,N-1}, v_{t,N}]^T$는 도로 내 모든 차량의 절대 속력을 의미하며, $\boldsymbol{p_t} = [p_{t,1}, p_{t,2}, \cdots, p_{t,N-1}, p_{t,N}]^T$는 차량의 절대 위치를 의미한다. $\boldsymbol{k_t} = [k_{t,1}, k_{t,2}, \cdots, k_{t,N-1}, k_{t,N}]^T$는 각 차량이 위치한 도로의 차선 번호이며, $\boldsymbol{n_t} = [n_{t,1}, n_{t,2}, \cdots, n_{t,N-1}, n_{t,N}]^T$는 각 차량이 위치한 도로의 전체 차선 개수를 의미한다.

3.2 병목구간 통과를 위한 Partially Observable Markov Decision Process

본 연구에서 개체는 상태 정보에 대한 부분적인 관측 정보를 통해 행동을 결정한다. 구체적으로, 개체는 본인이 위치한 절대 위치 $p_{t,N}$ 기준 전, 후방 W 범위 내의 차량을 관측할 수 있다. 또한 개체는 본인이 위치한 차선 번호 $k_{t,N}$를 포함하여 최대 H개의 차선을 관측할 수 있다. 그림 2는 개체가 관측할 수 있는 차량을 나타낸다. 즉, 관측 가능 차량 집합 C_{obs}은 도로 내 임의의 차량 $c_i \in C$ 중 아래의 조건을 모두 만족하는 차량을 의미한다.

$$\begin{cases} c_i \neq c_N \\ |\triangle p_{t,i}| \leq W \\ |\triangle k_{t,i}| \leq \dfrac{H-1}{2} \end{cases}$$

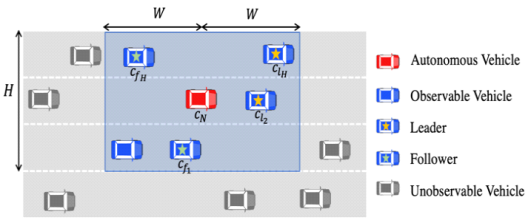


그림 2. 관측 가능 차량 정의
Fig. 2. The definition of observable vehicles

여기서 $\triangle(\cdot)$는 관측된 차량 C_{obs}와 개체 사이의 상대적인 상태정보를 의미한다. 즉, $\triangle p_{t,i}$와 $\triangle k_{t,i}$은 각각 개체와의 상대 속도 및 차선번호의 차이를 의미하며, 각각은 다음과 같이 정의한다.

$$\begin{aligned} \triangle p_t &= p_{t,i} - p_{t,N} \\ \triangle k_t &= k_{t,i} - k_{t,N} \end{aligned} \tag{1}$$

Observation : 관측 정보 $o_t \in \mathbb{R}^{5H+5}$는 개체 c_N 및 관측 가능 차량 C_{obs}의 상태정보를 기반으로 정의되며 다음과 같이 정의한다.

$$o_t = [\boldsymbol{\Delta v_t}^T, \boldsymbol{\Delta p_t}^T, \boldsymbol{\Delta \rho_t}^T, n_{t,W}, v_{t,N}, p_{t,N}, k_{t,N}, n_{t,N}]^T$$

이때, $\boldsymbol{\Delta v_t} = [\triangle v_{t,l_1}, \cdots, \triangle v_{t,l_H}, \triangle v_{t,f_1}, \cdots, \triangle v_{t,f_H}]^T$는 차선별 leader/follower 차량과의 상대 속도, $\boldsymbol{\Delta p_t} = [\triangle p_{t,l_1}, \cdots, \triangle p_{t,l_H}, \triangle p_{t,f_1}, \cdots, \triangle p_{t,f_H}]^T$는 상대 거리를 의미한다. 이때, 임의의 관측 가능 차선 $h \in [1, \cdots, H]$에서의 leader 차량 c_{l_h}은 차선별 관측된 전방 차량 집합 C_{obs,l_h} 중 자율주행차량과의 상대거리 절댓값이 최소인 차량을 의미한다. 이어서, follower 차량 c_{f_h}은 차선별 관측된 후방차량 집합 C_{obs,f_h} 중 자율주행차량과의 상대 거리 절댓값이 최소인 차량을 의미한다. $\boldsymbol{\rho_t}$는 전방 차선별 차량 밀도로 관측가능 거리 W 대비 관측된 차량이 차지하고 있는 도로 비율을 의미한다. 따라서 특정 h번째 차선의 차량 밀도 $\rho_{t,h}$는 해당 차선에 관측된 차량 대수 $|C_{obs,l_h}|$에 대해서 다음을 만족한다.

$$\rho_{t,h} \propto \frac{|C_{obs,l_h}|}{W} \tag{2}$$

즉, $\rho_{t,h}$는 관측가능 거리 W 대비 h번째 차선에서 관측된 모든 차량이 차지하고 있는 도로 비율로 해석할

수 있다. $n_{t,W}$는 전방 관측 가능 거리 W 이후 도로의 전체 차선 개수를 의미한다.

Action : 개체는 관측 정보 W를 기반으로 행동 $a_t=\{a_{t,acc},a_{t,lc}\}$을 결정한다. 여기서 $a_{t,acc}$는 가속도 조절을 의미하며, $a_{t,lc}$는 차선변경을 의미한다. 가속도 조절 $a_{t,acc}\in[a_{min},a_{max}]$은 최소 가속도 값 a_{min}과 최대 가속도 값 a_{max}의 범위 내에서 연속적인 값을 선택한다. 차선 변경 조절 $a_{t,lc}\in\{-1,0,1\}$은 이산적인 행동 공간에서 정의되며 각각의 값은 개체가 차선 변경을 수행할 방향을 의미하며 다음과 같이 정의한다.

$$\begin{cases} Left & a_{t,lc}=1 \\ Straight & a_{t,lc}=0 \\ Right & a_{t,lc}=-1 \end{cases}$$

Reward : 보상 r_t은 현재 상태 s_t, 현재 행동 a_t 그리고 다음 상태 s_{t+1}에 대한 함수 형태 $\mathrm{R}(s_t,a_t,s_{t+1})$로 존재하며 다음과 같다.

$$\mathrm{R}(s_t,a_t,s_{t+1})=\eta_1R_1+\eta_2R_2+\eta_3R_3+\eta_4R_4 \quad (3)$$

여기서 $R_{r\in\{1,2,3,4\}}$은 보상 항 또는 처벌 항을 의미하며, η_r는 R_r에 대한 계수를 의미한다.

R_1는 개체의 t 시점 행동 a_t에 따른 다음 시점 속력 $v_{t+1,N}$과 관련된 보상 함수로 목표 속력 v^* 및 제한 속력 v_{limit}에 대해 다음과 같다.

$$R_1=\begin{cases} \dfrac{v_{t+1,N}}{v^*} & v_{t+1,N}\le v^* \\ \dfrac{v_{limit}-v_{t+1,N}}{v_{limit}-v^*} & v_{t+1,N}>v^* \end{cases} \quad (4)$$

수식 (4)에 따라 개체는 목표 속력 v^*과 가깝게 주행할수록 최대의 보상을 획득하며, v^*을 초과한 속력을 낼 경우 보상은 선형적으로 감소하게 된다. 또한, 개체가 제한 속력 v_{limit} 이상의 속력으로 주행할 경우 음의 보상 값을 획득한다.

R_2는 성공적인 차선 변경을 위한 보상함수로 개체가 차선 변경을 수행한 경우 즉, $|a_{t,lc}|=1$일 때 적용되며 다음과 같이 정의한다.

$$R_2=|a_{t,lc}|(\triangle p_{t+1,l}-\triangle p_{t,l}-\delta_{lc}) \quad (5)$$

여기서, $\triangle p_{t,l}$와 $\triangle p_{t+1,l}$은 개체의 차선 변경 수행 전, 후 동일 차선 leader와의 상대 거리를 의미하며, $\delta_{lc}\in[0,W]$는 개체의 성공적인 차선 변경 기준을 결정하는 임계값(threshold)을 의미한다(그림 3). 즉, 개체가 δ_{lc} 이상의 상대 거리 이득($\triangle p_{t+1,l}-\triangle p_{t,l}\ge\delta_{lc}$)을 얻는 차선 변경을 수행했을 경우를 성공적인 차선 변경으로 해석하며, $\triangle p_{t+1,l}-\triangle p_{t,l}<\delta_{lc}$인 경우를 의미 없는 차선 변경으로 해석한다. 이를 통해 개체의 의미 없는 차선 변경은 약화하면서, 성공적인 차선 변경은 강화할 수 있다.

R_3은 동일 차선 follower의 안전거리 s^*를 침범하는 행동을 약화하기 위한 보상 함수로 개체가 차선 변경을 수행한 경우 즉, $|a_{t,lc}|=1$일 때 적용되며 다음과 같이 정의한다.

$$R_3=|a_{t,lc}|\times\min\left[0,1-\left(\frac{s^*}{\triangle p_{t+1,f}}\right)^2\right] \quad (6)$$

여기서, s^*와 $\triangle p_{t+1,f}$는 동일 차선 follower에 관한 요소로 각각은 해당 follower의 안전거리 및 개체와의 상대 거리를 의미한다. 여기서 안전거리 s^*는 최소 안전거리 s_0 및 사고 방지를 위한 최소 시간 s^*에 대해 다음과 같다.[23]

$$s^*=s_0+\max\left[0,v_{t,f}\cdot t^*+\frac{v_{t,f}(v_{t,f}-v_{t,N})}{2\sqrt{a_{max}\cdot|a_{min}|}}\right] \quad (7)$$

즉, $\triangle p_{t+1,f}<s^*$인 경우를 follower의 안전거리를 침범한 행동으로 간주하여 페널티를 부여한다.

R_4는 개체가 수행 불가한 차선 변경 행동을 결정했을 때 부여하는 페널티이다. 예를 들어 차선이 하나인 도로에서 개체가 차선 변경 행동을 결정하는 경우 페널티를 부여한다.

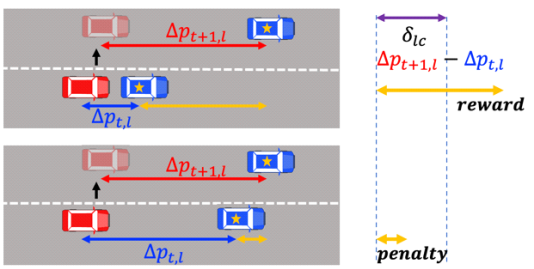


그림 3. 성공적인 차선 변경 임계값
Fig. 3. The threshold for a successful lane change

3.3 TD3 및 우선적 경험 재생 기반 정책학습

본 연구에서는 성공적인 병목 통과 정책 학습을 위해 액터-크리틱 기반의 심층강화학습 알고리즘인 TD3를 사용한다. TD3의 네트워크 구조는 정책 근사를 위한 액터 네트워크 π_ϕ와 Q 값을 근사하기 위한 두 개의 크리틱 네트워크 $Q_{\theta_1}, Q_{\theta_2}$ 및 각각에 대한 target 네트워크 $\pi_{\phi'}, Q_{\theta_1'}, Q_{\theta_2'}$로 구성된다. TD3는 off-policy 기반의 알고리즘으로 경험 재생을 통해 경로정보 $<o_t, a_t, r_t, o_{t+1}>$의 재사용이 가능하다. 본 연구에서 고려하는 prioritized 경험 재생은 버퍼에 저장한 정보의 우선순위 p_d에 따라 학습에 사용될 확률 $P(d)$을 결정하며, $P(d)$는 다음과 같이 정의한다.

$$P(d) = \frac{p_d^{w_{pr}}}{\sum_{d=1}^{D} p_d^{w_{pr}}} \tag{8}$$

여기서 D는 버퍼 크기를 의미하며, w_{pr}는 우선순위 규칙을 적용할 정도를 결정하는 계수이다. 이때, $w_{pr}=0$인 경우에는 random 경험 재생과 동일하게 작동한다. 우선순위 p_d는 해당 정보의 TD-error인 δ_d를 통해 결정되며 다음과 같이 정의한다.

$$p_d = |\delta_d| + \epsilon = \left| y_d - Q_{\theta_i}(o_d, a_d) \right| + \epsilon \tag{9}$$

여기서 ϵ은 $\sum_{d=1}^{D} |\delta_d|^{w_{pr}} = 0$일 때 $P(d)$의 분모가 0이 되는 것을 방지하기 위한 충분히 작은 상수이다. y_d는 TD3에 기반한 TD-target으로 다음과 같이 정의한다.

$$\begin{aligned} y_d &= r_d + \gamma \min_{i=1,2} Q_{\theta_i'}(o_{d+1}, \pi_{\phi'}(o_{d+1}) + \epsilon_{TD3}) \\ where \;\; \epsilon_{TD3} &\sim clip(N(0, \sigma_{TD3}^2), -\mu, \mu) \end{aligned} \tag{10}$$

이때, ϵ_{TD3}는 정책 평활화(smoothing)를 위한 노이즈로 $\pm\mu$ 값으로 클리핑된 평균이 0, 분산이 σ_{TD3}^2인 정규분포 $N(0, \sigma_{TD3}^2)$에서 샘플링한다. 수식 (10)을 통해 TD3의 TD-target y_d은 두 개의 크리틱 네트워크 출력 중 작은 Q 값을 통해 결정되는 것을 확인할 수 있다. 이는 TD3의 특성으로, 작은 Q 값을 target 값으로 설정함으로써 과대 추정(overestimation) 문제를 완화한다.

Prioritized 경험 재생 방식에서는 우선순위가 갱신될 때 변경되는 샘플링 확률분포를 정정하기 위해 Importance Sampling (IS) 가중치를 고려한다. IS 가중치 w_d는 실제 샘플링 확률 분포 $P_{origin}(d)$ 와 학습에 사용되는 데이터의 샘플링 확률 분포 $P_{train}(d)$가 다를 경우 분포 차이를 보상하기 위한 가중치로 다음과 같이 정의된다.

$$w_d = \frac{P_{origin}(d)}{P_{train}(d)} \tag{11}$$

Prioritized 방식에서 $P_{origin}(d)$는 무작위 선택 즉, $w_{pr}=0$인 경우이며, $P_{train}(d)$는 $P(d)$이므로 선택된 각 정보에 대한 IS 가중치 w_d는 다음과 같이 정의할 수 있다.

$$w_d = \left(\frac{1}{D} \cdot \frac{1}{P(d)} \right)^{w_{is}} \tag{12}$$

여기서 w_{is}는 IS 가중치를 적용하는 정도를 결정하는 계수를 의미한다.

액터 및 크리틱 네트워크의 업데이트는 샘플링 확률 분포 $P(d)$를 통해 배치 단위로 샘플링된 B개의 경로 정보를 기반으로 이루어진다. 이때, TD3는 액터 네트워크의 업데이트 주기를 크리틱 네트워크의 비해 u_d 만큼 지연한다. 이는 여러 번의 업데이트를 통해 보다 정확하게 근사된 Q 값을 액터 네트워크 정책 평가에 사용하기 위한 목적이다.

크리틱 네트워크의 업데이트는 크리틱 목적함수 $J(\theta_i)$에 대해 경사하강법을 적용하는 방식으로 이루어지며, 학습률 α_{critic}에 대해서 다음과 같이 정의한다.

$$\begin{aligned} \theta_i &\leftarrow \theta_i - \alpha_{critic} \nabla_{\theta_i} J(\theta_i) \\ where \;\; \nabla_{\theta_i} J(\theta_i) &= \frac{\partial J}{\partial Q_{\theta_i}} \frac{\partial Q_{\theta_i}}{\partial \theta_i} \end{aligned} \tag{13}$$

이때, $J(\theta_i)$는 샘플링 확률분포 정정을 위한 IS 가중치가 고려된 TD-error의 평균 절대오차 형태로 정의되며, 배치 내 경로 정보 j에 대해 다음과 같이 나타낸다.

$$J(\theta_i) = \frac{1}{B} \sum_{j=1}^{B} \left(y_j - Q_{\theta_i}(o_j, a_j) \right)^2 \tag{14}$$

액터 네트워크의 업데이트는 크리틱 네트워크보다 u_d만큼 지연된 시점에서 액터의 목적 함수 $J(\phi)$에 대해 경사 상승법을 적용함으로써 이루어지며, 학습률 α_{actor}에 대해서 다음과 같이 정의한다.

$$\phi \leftarrow \phi - \alpha_{actor} \nabla_{\phi} J(\phi)$$
$$where \quad \nabla_{\phi} J(\phi) = \frac{\partial J}{\partial \pi_{\phi}} \frac{\partial \pi_{\phi}}{\partial \phi} \tag{15}$$

이때, 액터의 목적함수 $J(\phi)$는 Q_{θ_1} 네트워크의 Q 값을 기반으로 결정되며, 배치 내 경로 정보 j에 대해 다음과 같이 정의한다.

$$J(\phi) = \frac{1}{B} \sum_{j=1}^{B} Q_{\theta_1}(o_j, \pi_{\phi}(o_j)) \tag{16}$$

각 네트워크의 target 네트워크에 대한 업데이트는 액터 네트워크와 동일한 시점에 수행된다. 이때, TD3에서는 soft update 방식을 사용한다. Soft update는 target 네트워크를 업데이트하는 과정에서 네트워크 파라미터가 급격하게 변화하는 것을 방지하기 위한 방법으로 $\tau \in [0,1]$를 가중치로 설정하여 간접적으로 target 네트워크를 갱신한다.

$$\theta_i' \leftarrow \tau \theta_i + (1-\tau)\theta_i'$$
$$\phi' \leftarrow \tau \phi + (1-\tau)\phi' \tag{17}$$

자세한 구동 방식은 알고리즘 1을 통해 확인할 수 있다. 먼저 정책 학습 과정에 사용되는 hyperparameter 설정을 수행한다. 이어서, 학습 전 replay buffer $\mathcal{H}$ 및 네트워크를 초기화한다. 개체는 T_{eps} 만큼의 길이를 가지는 에피소드를 E번 만큼 반복함으로써 정책을 학습하며, 에피소드가 시작되는 시점에 최초 상태 정보 s_1가 초기화한다. 개체는 매 시점 t에서의 경로 정보를 버퍼 $\mathcal{H}$에 저장한다. 이후 앞에서 기술한 prioritized 경험 재생 방식에 따라 액터 및 크리틱 네트워크 그리고 target 네트워크의 업데이트를 진행한다. 구체적인 hyperparameter 설정은 Appendix A2에서 확인할 수 있다.

Ⅳ. 성공적인 병목 통과 성능 측정 및 분석

본 절에서는 자율주행차량의 정책학습 결과 분석 및 성능 평가를 수행한다. 우선적으로, 도로 환경 구축을 위한 시뮬레이터과 학습 환경설정을 살펴본 뒤, 학습된 차량의 성능 분석을 위한 사분범위(Interquartile Range; *IQR*) 해석법에 대해 소개한다. 이후, prioritized 및 random 경험 재생 방식을 통해 학습된 자율주행 차량의 주행 데이터 분석을 통해 성능 비교를 수행한다.

Algorithm 1 TD3-based training algorithm with prioritized experience replay

Require: priority coefficient w_{pr}, importance sampling coefficient w_{is}, policy delay u_d, the number of episodes E, and the number of training time step T_{eps}

Initialization: replay buffer $\mathcal{H}$, critic networks Q_{θ_1}, Q_{θ_2}, actor network π_{ϕ}, target networks: $Q_{\theta_1'} \leftarrow Q_{\theta_1}$, $Q_{\theta_2'} \leftarrow Q_{\theta_2}$, $\pi_{\phi'} \leftarrow \pi_{\phi}$

for episode $e = 1$ to E **do**
  Reset state s_1
  for $t = 1$ to T_{eps} **do**
    Get observation o_t from state s_t and select action $a_t \sim \pi_{\phi}(o_t)$
    Observe next observation o_{t+1} and observe R_t
    Store $(o_t, a_t, \mathrm{R}_t, o_{t+1})$ in $\mathcal{H}$ with maximal priority
    Sample B trajectory from $\mathcal{H}$ by $P(d)$:
    $P(d) = p_d^{w_{pr}} / \sum_{d=1}^{D} p_d^{w_{pr}}$
    Compute TD-target y based on (10)
    Update critics θ_i by gradient descent:

$$J(\theta_i) = B^{-1} \sum_{j=1}^{D} w_j \cdot \delta_j^2,$$
$$where \quad w_j = (B \cdot P(j))^{-w_{is}}, \quad \delta_j = y_j - Q_{\theta_i}(o_j, a_j)$$

    Update trajectory priority $p_j \leftarrow |\delta_j|$
    if t mod u_d then
      Update actor ϕ by gradient ascent:
      $J(\phi) = B^{-1} \sum_{j=1}^{B} Q_{\theta_1}(o_j, \pi_{\phi}(o_j))$
      Update target networks:
      $\theta_i' \leftarrow \tau \theta_i + (1-\tau)\theta_i'$
      $\phi' \leftarrow \tau \phi + (1-\tau)\phi'$
    end if
  end for
end for

알고리즘 1. Prioritized 경험 재생 방식을 통한 TD3 기반 자율주행 정책 학습 알고리즘
Algorithm. 1. TD3-based training algorithm with prioritized experience replay

4.1 시뮬레이터 및 학습 환경 설정

본 연구에서는 도로 환경 설정 및 실험을 위해 교통 제어 시뮬레이터 SUMO[24] 기반의 FLOW[25]를 사용하며, 도로 환경은 그림 2와 같이 병목구간이 포함된 도로를 고려한다. 이때, 도로 내에는 자율주행 차량과 일반차량이 공존하는 환경을 고려한다. 또한, 병목 구간에서 개체가 충분한 정체현상을 경험할 수 있도록 고려하였다. 이를 위해 병목지점의 개수 $Y = 2$

로 설정하였으며 도로는 최대 4개, 최소 2개의 차선을 갖도록 설정하였다. 전체 차량 대수 $N=32$이며 도로 길이 $L=465m$로 설정함으로써 충분한 차량정체가 발생하도록 설정하였다. 차량 설정으로는 도로 내 모든 차량 c_i의 목표 속력 $v^*=12.5m/s$이며, 자율주행차량 c_N의 제한 속력 $v_{limit}=15m/s$로 설정하였다. 자율주행차량의 가속도 범위는 최대 가속도 $a_{max}=1m/s^2$, 최소 가속도 $a_{min}=-1m/s^2$로 설정하였다. 마지막으로, 에피소드 실행 시간 T_{full}은 $3900t_s$이며, 이는 학습 안정화를 위한 시동(warm-up) 시간 $T_w=900t_s$과 정책 학습을 시간 $T_{eps}=3000t_s$로 구성된다. 여기서 $1t_s$는 $0.1s$로 설정하였다. 구체적인 시뮬레이터 hyperparameter 설정은 Appendix A3 에서 확인할 수 있다.

시뮬레이션 내 차량은 일반차량과, 자율주행 차량으로 구성된다. 이때, 일반차량은 차선 변경의 수행 없이 가속도 조절만을 수행한다[1]. 일반차량의 가속도 조절은 Intelligence Driving Module (IDM) 컨트롤러에 의해 수행되며, 다음과 같은 수식에 기반한다[23].

$$a_{t,i}=a_{max}\left(1-\left(\frac{v_{t,i}}{v^*}\right)^{\psi}-\left(\frac{s^*}{\triangle p_{t,l}}\right)^2\right)+\epsilon_{IDM} \quad (18)$$
$$where \quad \epsilon_{IDM}\sim N(0,\sigma^2_{IDM})$$

여기서 ϵ_{IDM}은 평균이 0, 분산이 σ^2_{IDM}인 정규 분포에서 샘플링되는 잡음을 의미한다. 해당 잡음은 IDM 컨트롤러에 의한 가속도 조절에 불확실성으로 작용하기 때문에 보다 현실적인 도로 환경을 고려할 수 있다. 자율주행차량은 학습된 정책 네트워크 π_ϕ의 출력값을 통해 가속도 조절 $a_{t,acc}$ 및 차선 변경 $a_{t,lc}$을 수행한다.

4.2 성능평가 지표 및 해석

본 연구에서는 자율주행 차량의 주행 성능을 평가하기 위해 IQR 활용한 데이터 분석을 수행한다(그림 4). IQR 해석은 전체 데이터를 4개의 동등한 비율로 분할하는 사분위수(quartile) 값인 $Q1$, $Q2$, $Q3$에 기반하며, 이상치(outliers)를 제외한 데이터 구간에 대한 분석이 가능하기 때문에 정규분포를 따르지 않는 데이터 분석에도 적용할 수 있다. 여기서 $Q1$과 $Q3$는 각각 전체 데이터에서 25%, 75%에 해당하는 값

1) 단, 차선이 증가하는 구간에 한해서는 일반차량의 차선 변경을 허용한다.

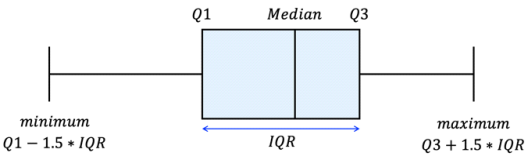


그림 4. 사분 범위 정의
Fig. 4. The definition of interquartile range

을 의미하며, $Q2$는 전체 데이터의 50%에 해당하는 중앙값(median)을 의미한다. IQR은 $Q1$과 $Q3$ 사이의 데이터 범위를 의미하며, 다음과 같이 정의한다.

$$IQR=Q3-Q1 \quad (19)$$

IQR 값은 이상치를 제외한 정상 범위 내의 데이터를 판별하는 기준으로 고려된다. 이때, 정상 범 위 내 데이터의 최소 및 최댓값은 다음과 같다.

$$maximum=Q3+1.5*IQR$$
$$minimum=Q1-1.5*IQR \quad (20)$$

4.3 자율주행 차량의 주행 성능 평가

본 연구에서는 경험 재생 방식의 차이가 병목 구간 통과 자율주행 정책에 미치는 영향을 분석 우선적으로, 경험 재생 방식별 학습 결과를 에피소드당 누적 보상 그래프를 통해 비교한 뒤 prioritized 및 random 경험 재생 기반 차량의 차선 변경 경향성을 IQR을 통해 분석한다. 이후, 차선 변경 경향성의 차이가 주행 성능에 미치는 영향을 확인하기 위해 복잡도가 높은 도로 환경에서의 경험 재생 방식별 주행 성능을 정성적/정량적으로 평가한다.

4.3.1 경험 재생 방식에 따른 학습 결과

그림 5는 에피소드가 증가함에 따른 경험 재생 방식별 평균 누적 보상 그래프이다. 해당 결과는 각각의

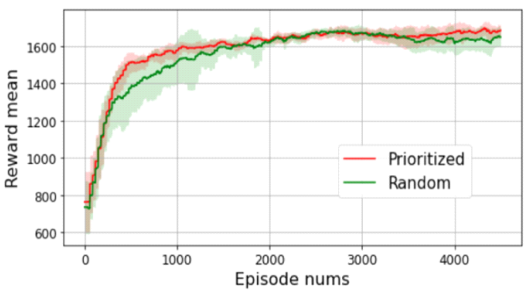


그림 5. 경험 재생 방식 별 학습 에피소드에 따른 누적 보상 평균
Fig. 5. The accumulative mean reward based on experience replay types

경험 재생 방식에서 10개의 랜덤 시드에 대해 측정되었다. 그래프에서 실선은 해당 에피소드 구간에서 랜덤 시드간 평균 누적 보상을 의미하며 음영은 표준 편차를 의미한다. 해당 그래프를 통해 두 가지 방식 모두 1600 ~ 1660 사이에서 수렴하는 것을 확인할 수 있다. 이때, 전반적으로 prioritized 방식의 분산이 random 방식보다 낮은 것을 확인할 수 있다. 이는 학습 단계에서 경로 정보의 우선순위를 고려하는 것이 랜덤 시드에 대한 정책 일관성이 높다는 것을 의미한다.

4.3.2 경험 재생 방식별 차선 변경 경향성 분석

경험 재생 방식의 차이는 학습된 개체의 차선 변경 경향성의 차이를 야기한다. 본 절에서는 경험 재생 방식에 따른 차선 변경 경향성을 분석하기 위해, 차선 변경 전 자율주행 차량과 선두 차량 간의 상대 거리 $|a_{t,lc}|\triangle p_{t,l}$및 차선 변경 후 자율주행차량의 속력 $|a_{t,lc}|v_{t+1,N}$을 분석한다.

우선적으로, $|a_{t,lc}|\triangle p_{t,l}$에 대한 정성적인 분석은 그림 6을 통해 확인할 수 있다. 해당 그림은 $|a_{t,lc}|\triangle p_{t,l}$의 전체 데이터 분포에 대한 IQR 범위를 경험 재생 방식에 따라 나타낸 결과이다. 해당 결과를 통해 prioritized 방식의 개체가 random 방식의 개체보다 전반적으로 선제적인 차선 변경을 수행하는 것을 확인할 수 있다.

보다 자세한 수치는 표 1을 통해 확인할 수 있다. 해당 표를 통해 $|a_{t,lc}|\triangle p_{t,l}$에 대한 모든 사분위수가 prioritized 방식에서 높게 측정되는 것을 확인할 수 있다. 특히, $Q1$ 값의 경우 random 방식이 $6.98m$ 만큼 낮게 측정된 것을 확인할 수 있다. 이는 random 방식의 개체가 차선 변경 시 동일 차선 선두 차량과 상대적으로 가까운 지점에서 차선 변경을 수행하는 횟수가 빈번하다는 것을 의미한다.

동일 차선 선두 차량과 가까운 거리에서 차선을 변

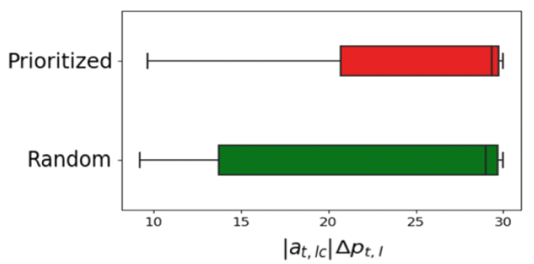


그림 6. 경험 재생 방식별 차선 변경 시 동일 차선 선두차량과의 상대거리 사분범위
Fig. 6. The interquartile range of relative position to the same lane leader at the time of lane changes

표 1. 차선 변경 시 선두 차량과의 상대 거리에 대한 정량적 분석
Table 1. The quantitative results of relative position to the leader at the time of lane changes

	$\|a_{t,lc}\|\triangle p_{t,l}\ (m)$	
Replay type	*Median*	$Q1-Q3$
Prioritized	**29.37**	**20.71 − 29.73**
Random	29.01	13.73 − 29.01

경하는 경우, 해당 차량은 선두 차량의 속력 감소에 비교적 큰 영향을 받게 된다. 이는 그림 7을 통해 확인할 수 있다. 해당 그림은 차선 변경 후 속력 $|a_{t,lc}|v_{t+1,N}$ 분포 및 IQR 범위를 경험 재생 방식별로 나타낸 결과이다. 그림 7의 분포 그래프를 통해 prioritized 방식의 속력 분포는 목표 목표속도인 $12.5m/s$ 주변에 밀집되어 있으며, random 방식의 경우에는 낮은 속력 구간에서 높은 분포를 보이는 것을 확인할 수 있다. 이때, $|a_{t,lc}|v_{t+1,N}$의 분포는 2개의 최빈값(mode)을 갖는 것을 확인할 수 있다. 따라서, IQR 관점의 해석을 위해 전체 분포 구간을 2개의 구간인 $Section\,A, B$로 나누어서 분석을 진행하였다[2]. 먼저, $Section\,A$ 구간의 IQR 박스를 통해 random 방식은 prioritized 방식에 비해 작은 속력 값 주위에 밀집되어 있는 것을 확인할 수 있다. 또한, $Section\,B$의 IQR 박스를 통해 prioritized 방식의 개체가 전반적으로 차선 변경 후 높은 속력을 유지하고 있는 것을 확인할 수 있다.

보다 자세한 분석은 표 2의 측정값을 통해 확인할 수 있다. 해당 표는 각 $Section$별 IQR 범위 및 데이터 분포 특성을 정량적으로 측정한 결과이다. 여기서 데이터 비율(data percentage)은 경험 재생 방식별 전체 데이터가 해당 $Section$에 존재하는 비율을 의미한다. 데이터 비율의 비교를 통해 prioritized 방식은 높은 속력 구간인 $Section\,B$에서 88.6%의 비율을 형성함에 비해 random 방식은 53.2%에 불과한 것을 확인할 수 있다. 이는 prioritized 방식이 random 방식보다 차선 변경 시 높은 속력을 유지한다는 것을 의미한다. 또한, $Section\,A$에서의 중앙값은 두 가지 경험 재생 방식 모두 $0.22m/s$인 것을 확인할 수 있는데, 이는 낮은 속력 구간에서 자율주행차량은 주로 정지에 가까운 상태에서의 차선을 변경한다는 것을 의미한다. 마지막으로, $Section\,B$에서의 모든 측정값에서

2) $Section\,A$와 B는 random 경험 재생 방식에서의 $|a_{t,lc}|v_{t+1,N}$ 분포 중 가장 작은 비율을 갖는 값을 기준으로 분리되었다.

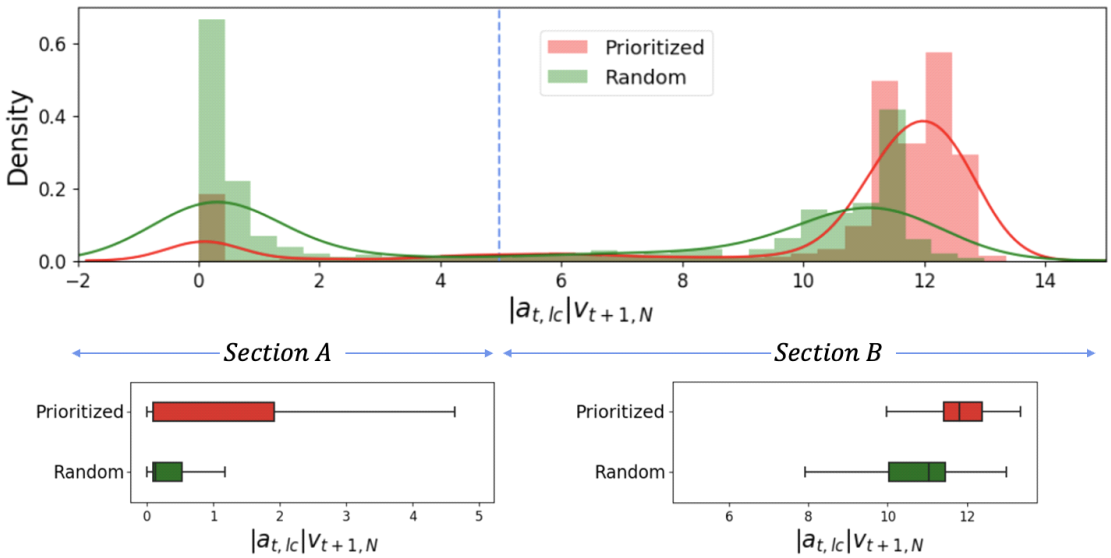


그림 7. 경험 재생 방식 별 차선 변경 시 자율주행차량의 속력 분포 및 사분범위
Fig. 7. The speed distribution and interquartile range of the autonomous vehicle after lane changes

표 2. 차선 변경 시 자율주행 차량 속력에 대한 정량적 분석
Table 2. The quantitative results of the autonomous vehicle speed after lane changes

Replay type	*Section A*				*Section B*			
	Data percentage	$\|a_{t,lc}\|v_{t+1,N}$ (m/s)			Data percentage	$\|a_{t,lc}\|v_{t+1,N}$ (m/s)		
		Mode	*Median*	$Q1-Q3$		*Mode*	*Median*	$Q1-Q3$
Prioritized	11.4%	0.22	0.09	0.09 − 1.91	**88.6%**	**12.25**	**11.80**	**11.41 − 12.38**
Random	46.8%	0.22	0.13	0.09 − 0.52	53.2%	11.48	11.03	10.34 − 11.45

prioritized 방식의 개체가 우수함을 확인할 수 있다. 이를 통해 학습 과정에서 경험 정보의 우선순위를 고려하여 학습된 개체는 선제적인 차선 변경을 수행하며 더 나아가 개체가 주행 중 높은 속력을 유지할 수 있음을 확인할 수 있다.

4.3.3 도로 복잡도 증가에 따른 주행 성능 비교

경험 재생 방식에 따른 주행 성능 차이는 도로 환경이 복잡해짐에 따라 증가한다. 이는 그림 8을 통해 확인할 수 있다. 해당 그래프는 도로 내 전체 차량 대수 N이 증가함에 따른 prioritized 및 random 기반 개체의 전 구간 평균 속력 차이를 나타낸 결과이다. 이때, N이 증가할수록 차량 정체 현상은 심화되며 이에 따라 더욱 복잡한 도로 환경이 조성된다. 그래프를 통해 도로가 복잡해질수록 prioritized 방식의 개체가 더 빠른 주행을 수행함을 확인할 수 있다. 이는 학습 과정에서 경로 정보의 우선순위를 고려한 개체가 복잡한 도로에서 우수한 주행을 수행한다는 것을 의미한다.

도로의 복잡도가 가장 높은 환경($N=32$)에서의 구체적인 성능 비교는 그림 9의 (a)를 통해 확인할 수 있다. 해당 그래프는 도로 위치에 따른 차량의 속력 평균을 경험 재생 방식에 따라 나타낸 결과이다. 그래프에서 $120m \sim 270m$ 사이에 위치한 혼잡지역(congestion area)은 병목 정체 현상으로 인해 차량 밀도가 높은 구간을 의미한다. 그래프를 통해 차선 변경

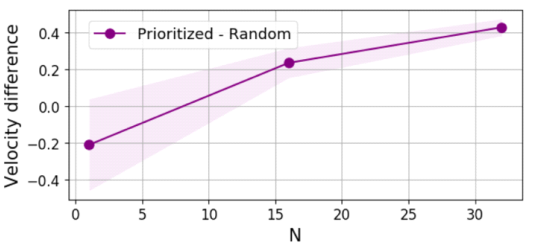


그림 8. 도로 내 전체 차량 대수에 따른 전 구간 평균 속력 차이
Fig. 8. The average speed difference between the total number of vehicles on the road

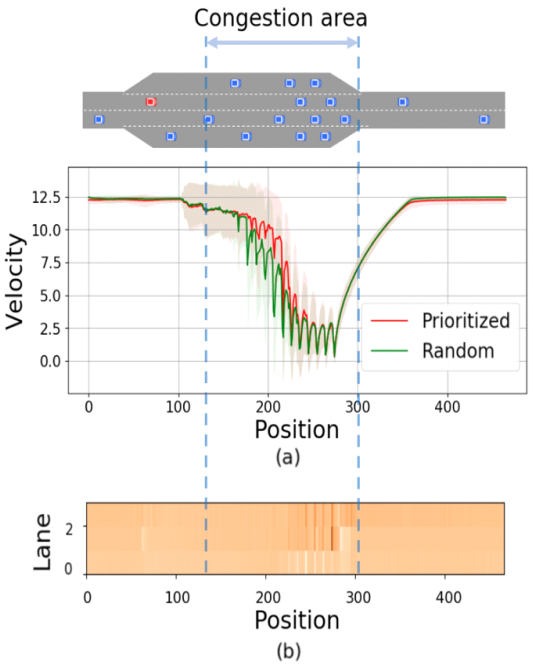


그림 9. 경험 재생 방식에 따른 도로 구간별 비교 (a) 평균 속력, (b)학습과정의 재생경험 빈도수 차이
Fig. 9. Comparison between replay types per road section (a) average speed, (b) the differences in data replay frequency in training

이 가장 많이 발생하는 병목 구간에서 prioritized 기반의 개체가 random 방식의 개체보다 높은 속력을 유지 하는 것을 확인할 수 있다. 이는 우선순위를 고려한 개체가 선제적 차선 변경을 통해 높은 속력을 유지하면서 병목 구간을 통과하는 것으로 해석할 수 있다.

병목구간에서 prioritized 방식이 random 방식보다 우수한 주행 특성을 보이는 것은 그림 9의 (b)를 통해 설명할 수 있다. 해당 히트맵은 TD3 학습 과정에서 각 경험 재생 방식별 사용된 데이터의 빈도수 차이를 도로 구간별로 나타낸 결과이다. 해당 히트맵에서 어두운 구간은 prioritized 방식에서 자주 사용된 데이터를 의미하며, 밝은 구간일수록 random 방식에서 높은 빈도수로 사용되었음을 의미한다. 해당 히트맵을 통해 정체가 빈번히 발생하는 병목 구간에서 학습 데이터 빈도수 차이가 가장 크게 발생하는 것을 확인할 수 있다. 결과적으로, prioritized 기반의 개체는 중요도가 높은 병목 구간 데이터를 학습에 자주 사용하기 때문에 random 방식의 개체에 비해 전반적인 병목 도로 구간에서 최적의 행동을 수행하도록 학습된다. 본 연구에서 개체의 낮은 속력은 보상 감소에 직접적인 영향을 미치기 때문에 prioritized 방식의 개체는 높은 보상을 유지하기 위해 선제적인 차선 변경을 통해 속력 감소를 최소화하도록 학습된다.

V. 결 론

본 연구에서는 차량 정체가 빈번히 발생하는 병목 구간을 성공적으로 통과하기 위한 POMDP를 제안하였다. 또한, TD3 알고리즘을 통한 학습 과정에서 경험 재생 방식의 차이가 학습된 개체의 주행 성능에 유의미한 차이를 도출한다는 것을 확인하였다. 이를 위해 random 및 prioritized 경험 재생 기반 개체의 주행 데이터에 대해 *IQR* 분석을 수행하였으며, 결과적으로 prioritized 경험 재생 기반의 개체가 random 방식의 개체보다 차선 변경을 선제적으로 수행하는 것을 보였다. 또한, 선제적 차선 변경 특성으로 인해 prioritized 기반의 개체가 높은 속력의 주행을 유지할 수 있는 것을 확인하였다. 마지막으로, 도로 내 복잡도 증가에 따른 경험 재생 방식별 주행 성능 비교를 통해 prioritized 기반의 개체가 복잡한 도로에 더욱 강건한 주행 성능을 보이는 것을 확인하였다.

References

[1] SAE, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (2021), Retrieved Nov. 30, 2022, from https://saemobilus.sae.org/content/j3016_202104/. (https://doi.org/10.4271/J3016_202104)

[2] N. Kim, Y. Eom, and M. Jung, *The Ecosystem of Electric Vehicles and Autonomous Driving Based on Global M&A* (2022), Retrieved Apr. 11, 2023, from https://assets.kpmg.com/content/dam/kpmg/kr/pdf/2022/issue-monitor/kr-im-automotive-industry-ma-20220719.pdf.

[3] S. Fugimoto, et al., "Addressing function approximation error in actor-critic methods," *ICML*, Stockholm, Sweden, Jul. 2018.

[4] T. Schaul, et al., "Prioritized experience replay," *ICLR*, San Juan, Puerto Rico, May 2016.

[5] A. Nagabandi and G. Kagn, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," *ICRA*, Brisbane, Australia, May 2018.

(https://doi.org/10.1109/ICRA.2018.8463189)

[6] S. Racanière, et al., "Imagination-augmented agents for deep reinforcement learning," *NIPS*, California, USA, Dec. 2017.

[7] V. Mnih, et al., "Playing atari with deep reinforcement learning," *NIPS Wkshp.*, Nevada, USA, Dec. 2013.

[8] H. Van Hasselt, et al., "Deep reinforcement learning with double Q-leaning," *AAAI*, Arizona, USA, Feb. 2016. (https://doi.org/10.1609/aaai.v30i1.10295)

[9] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, no. 3-4, pp. 229-256, May 1992. (https://doi.org/10.1007/BF00992696)

[10] N. Metropolis, et al., "Equations of state calculations by fast computing machines," *The J. Chem. Phys.*, vol. 21, no. 6, pp. 1087-1092, Jun. 1953. (https://doi.org/10.1063/1.1699114)

[11] T. Lillicrap, et al., "Continuous control with deep reinforcement learning," *ICLR*, San Juan, Puerto Rico, May 2016.

[12] T. Haarnoja, et al., "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *ICML*, Stockholm, Sweden, Jul. 2018.

[13] B. R. Kiran, et al., "Deep reinforcement. learning for autonomous driving: A survey," *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 6, pp. 4909-4926, Feb. 2021. (https://doi.org/10.1109/TITS.2021.3054625)

[14] D. Lee and M. Kwon, "Combating stop-and-go wave problem at a ring road using deep reinforcement learning based autonomous vehicles," *J. KICS*, vol. 46, no. 10, pp. 1667-1682, Oct. 2021. (https://doi.org/10.7840/kics.2021.46.10.1667)

[15] Y. Song, et al., "Autonomous overtaking in gran turismo sport using curriculum reinforcement learning," *ICRA*, Xi'an, China, May 2021. (https://doi.org/10.1109/ICRA48506.2021.9561049)

[16] P. Wang, et al., "A reinforcement learning based approach for automated lane change maneuvers," *IVS*, Changshu, China, Jun. 2018. (https://doi.org/10.1109/IVS.2018.8500556)

[17] D. Lee and M. Kwon, "ADAS-RL: Safety learning approach for stable autonomous driving," *ICT Express*, vol. 8, no. 3, pp. 479-483, Sep. 2022. (https://doi.org/10.1016/j.icte.2022.05.004)

[18] Z. Li, et al., "Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks," *IEEE Trans. Intell. Transport. Syst.*, vol. 18, no. 11, pp. 3204-3217, Jun. 2017. (https://doi.org/10.1109/TITS.2017.2687620)

[19] C. Wang, et al., "A new solution for freeway congestion: Cooperative speed limit control using distributed reinforcement learning," *IEEE Access*, vol. 7, pp. 41947-41957, Mar. 2019. (https://doi.org/10.1109/ACCESS.2019.2904619)

[20] F. Belletti, et al., "Expert level control of ramp metering based on multi-task deep reinforcement learning," *IEEE Trans. Intell. Transport. Syst.,* vol. 19, no. 4, pp. 1198-1207, Aug. 2018. (https://doi.org/10.1109/TITS.2017.2725912)

[21] C. Wang, et al., "Integrated traffic control for freeway recurrent bottleneck based on deep reinforcement learning," *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 9, pp. 15522-15535, Jan. 2022. (https://doi.org/10.1109/TITS.2022.3141730)

[22] E. Vinitsky, et al., "Lagrangian control through deep-rl: Applications to bottleneck decongestion," *ITSC*, Hawaii, USA, Nov. 2018. (https://doi.org/10.1109/ITSC.2018.8569615)

[23] M. Treiber, et al., "Congested traffic states in empirical observations and microscopic simulation," *Physical review E*, vol. 62, no. 2, pp. 1805-1824, Aug. 2000. (https://doi.org/10.1103/PhysRevE.62.1805)

[24] P. A. Lopez, et al., "Microscopic traffic simulation using SUMO," *ITSC*, Auckland, New Zealand, Oct. 2018. (https://doi.org/10.1109/ITSC.2018.8569938)

[25] C. Wu, et al., "Flow: A modular learning framework for mixed autonomy traffic," *IEEE Trans. Robotics*, vol. 38, no. 2, pp. 1270-1286, Jul. 2021. (https://doi.org/10.1109/TRO.2021.3087314)

엄 찬 인 (Chanin Eom)

2022년 8월 : 숭실대학교 전자정보공학부 IT융합전공 학사

2022년 9월~현재 : 숭실대학교 정보통신공학과 석사과정

<관심분야> 강화학습, 인공지능, 자율주행

[ORCID:0009-0005-6340-6635]

이 동 수 (Dongsu Lee)

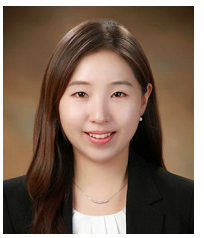
2022년 2월 : 숭실대학교 의생명시스템학부 빅데이터컴퓨팅융합전공 학사

2022년 3월~현재 : 숭실대학교 정보통신공학과 석박사통합과정

<관심분야> 강화학습, 계산신경과학, 자율주행

[ORCID:0000-0002-9238-4106]

권 민 혜 (Minhae Kwon)



2011년 8월 : 이화여자대학교 전자정보통신공학과 학사

2013년 8월 : 이화여자대학교 전자공학과 석사

2017년 8월 : 이화여자대학교 전자전기공학과 박사

2017년 9월~2018년 8월 : 이화여자대학교 전자전기공학과 박사 후 연구원

2018년 9월~2020년 2월 : 미국 Rice University, Electrical and Computer Engineering, Postdoctoral Researcher

2020년 3월~현재 : 숭실대학교 전자정보공학부 IT융합전공 조교수

<관심분야> 강화학습, 자율주행, 모바일네트워크, 연합학습, 계산신경과학

[ORCID:0000-0002-8807-3719]

Appendix

표 A1. 기호 정리 표
Table A1. Table of notation

Notation	Meaning	Notation	Meaning
S	state space	s_t	state
O	observation space	o_t	observation
A	action space	a_t	action
$a_{t,acc}$	acceleration action at t	$a_{t,lc}$	lane change action at t
a_{min}	minimum acceleration	a_{max}	maximum acceleration
R_t	reward function	γ	temporal discounted factor
M	a set of merge point	C	a set of vehicles
N	the number of vehicles on road	Y	the number of merge point on road
$c_i (i \neq N)$	human vehicles	c_N	an autonomous vehicle
W	visibility distance	H	the number of visible lanes
C_{obs}	set of observable vehicles	c_l	leader vehicle
c_f	follower vehicle	ρ	vehicle density on road ahead
v^*	desired velocity	v_{limit}	speed limit
δ_{lc}	successful lane change threshold	s^*	safety distance
L	road length	T_{full}	time steps of full episode
T_w	time steps for the warm-up stage	T_{eps}	time steps for the training stage
t_s	time step	ϵ_{IDM}	IDM controller noise
π_ϕ	actor network	Q_θ	critic network
ϕ'	target actor network parameter	θ'	target critic network parameter
B	batch size	E	number of training episodes
$\mathcal{D}$	replay buffer size	σ_{Td3}	standard deviation for action smoothing
μ	constant for noise clipping	u_d	delayed update period for actor network
α_{actor}	actor learning rate	α_{critic}	critic learning rate
ω_{pr}	the coefficient for PER	ω_{is}	the coefficient for IS

표 A2. 모델 하이퍼파라미터 설정
Table A2. Model hyperparameter settings

Parameter	Value
$Actor\ hidden$	$[64, 64, 64]$
$Critic\ hidden$	$[128, 128]$
α_{actor}	$\alpha_{actor} \in \{1e^{-3} \sim 1e^{-5}\}$
α_{critic}	$\alpha_{critic} \in \{5e^{-4} \sim 1e^{-6}\}$
γ	$\gamma \in \{0.78 \sim 0.99\}$
$Optimizer$	$Adam$
B	128
E	$E \in \{3000,\ 4500\}$
$\mathcal{D}$	$\mathcal{D} \in \{3 \times 10^6 \sim 10^8\}$
σ_{Td3}	0.2
μ	0.2
u_d	2
ω_{pr}	0.6
ω_{is}	0.4

표 A3. 시뮬레이터 설정
Table A3. Simulator settings

Parameter	Value
T_t	$3000t_s$
T_w	$900t_s$
N	32
W	$30m$
H	5
δ_{lc}	$5m$
v^*	$12.5m/s$
v_{limit}	$15m/s$
s_0	$2m$
t^*	$1s$
σ_{IDM}	0.2
$1t_s$	$0.1s$