ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 21-04-20
j.inst.Korean.electr.electron.eng.Vol.25,No.4,728~734,December 2021
https://doi.org/10.7471/ikeee.2021.25.4.728
145

종방향 주행성능향상을 위한 Latent SAC 강화학습 보상함수 설계

On the Reward Function of Latent SAC Reinforcement Learning to Improve Longitudinal Driving Performance

조 성 빈*, 정 한 유*★

Sung-Bean Jo*, Han-You Jeong*★

Abstract

In recent years, there has been a strong interest in the end-to-end autonomous driving based on deep reinforcement learning. In this paper, we present a reward function of latent SAC deep reinforcement learning to improve the longitudinal driving performance of an agent vehicle. While the existing reward function significantly degrades the driving safety and efficiency, the proposed reward function is shown to maintain an appropriate headway distance while avoiding the front vehicle collision.

요 약

최근 심층강화학습을 활용한 종단간 자율주행에 대한 관심이 크게 증가하고 있다. 본 논문에서는 차량의 종방향 주행 성능을 개선하는 잠재 SAC 기반 심층강화학습의 보상함수를 제시한다. 기존 강화학습 보상함수는 주행 안전성과 효율성이 크게 저하되는 반면 제시하는 보상함수는 전방 차량과의 충돌위험을 회피하면서 적절한 차간거리를 유지할 수 있음을 보인다.

Key words : Reinforcement learning, soft actor-critic, end-to-end learning, reward function, autonomous driving

Ⅰ. 서론

4차산업 혁명의 주요 기술 중 하나로 차량의 자율주행 기술에 대한 관심이 증가하고 있다. 자율주행은 차량에 장착된 다양한 센서데이터들을 이용해 주행환경을 인지하고 이를 바탕으로 스스로 판단하여 주행하는 시스템을 말한다. 일반적으로 자율주행 차량은 측위, 인지, 예측, 계획, 제어 파이프라인 형태의 고도로 모듈화된 구조로 설계된다. 이러한 모듈식 접근방식은 모듈 간 고정된 인터페이스로 인해 자율주행 기술의 최적화가 어렵고 유지보수 및 성능 개선의 복잡도가 크다.

최근 인공지능 기술이 발전함에 따라 특정 작업에 최적화된 기존 딥러닝 기술들을 다양한 작업들

* Dept. of Electrical Engineering, Pusan National University
★ Corresponding author
E-mail : hyjeong@pusan.ac.kr, Tel : +82-51-510-7332
※ Acknowledgment
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(2019R1I1A3A01060890).
Manuscript received Dec. 10, 2021; revised Dec. 27, 2021; accepted Dec. 29, 2021.

에 통합적으로 적용할 수 있도록 확장하는 연구들이 이루어지고 있다. 자율주행에서도 기존 모듈식 구성을 통합하여 센서데이터를 입력으로 직접 차량을 제어하는 종단간(End-to-End) 자율주행을 위한 강화학습에 대한 연구가 활발히 진행 중이다[1-5].

기존 강화학습 연구들에서 다양한 심층강화학습 신경망 구조들을 제시하였지만, 주행 성능을 개선하는 보상함수의 설계에 대한 연구는 아직 미흡한 실정이다[1-5]. 본 논문은 주행 안전성과 효율성을 향상하는 새로운 강화학습 보상함수를 제시한다. 자율주행 시뮬레이션을 위해 널리 사용하는 CARLA 시뮬레이터에 종방향 주행 환경을 설정하고 기존 보상함수와 새로운 보상함수를 기준으로 학습한 에이전트 차량의 주행 특성을 분석한다[6]. 종방향 주행 성능에 대한 체계적인 접근이 부족한 기존 보상함수는 전방 차량이 고속이로 이동하면 차간거리(Headway Distance)를 줄이지 못하는 문제점을 가지고 있다. 반면, 본 논문에서 제시하는 보상함수는 주행 속력에 상관없이 적절한 차간거리를 유지하면서도 전방 차량과의 교통사고를 회피할 수 있음을 보인다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 기존 강화학습 알고리즘과 보상함수 설계에 대한 관련 연구들을 소개한다. Ⅲ장에서는 종단간 자율주행을 위한 강화학습 에이전트와 보상함수를 제시한다. Ⅳ장에서는 CARLA 시뮬레이터를 활용한 성능평가를 수행하고, Ⅴ장에서는 본 논문의 결론을 제시한다.

Ⅱ. 관련 연구

1. 강화학습 알고리즘

강화학습은 기계학습의 한 분야로 순차적 행동결정문제의 모델링을 위해 마르코프 결정 프로세스(Markov Decision Process: MDP)를 사용하며, 시스템 상태, 에이전트의 행동, 행동에 대한 보상, 행동 결정의 정책을 정의한다. 강화학습은 주어진 환경과 상호작용하며 주어진 상태에서 보상을 최대화하는 에이전트 행동도출을 목표로 한다. 일반적으로 에이전트 보상을 최대화하는 방법에 따라 강화학습을 가치기반 알고리즘, 정책기반 알고리즘, 액터-크리틱(Actor-Critic) 기반 알고리즘으로 분류한다.

가치기반 알고리즘은 특정 상태에서 행동에 대한 가치인 Q 함수를 신경망으로 근사하여 최적의 Q 함수를 찾고, 탐욕 정책에 따라 최적의 행동을 수행한다. 대표적인 가치기반 알고리즘인 Deep-Q-Network(DQN)은 환경에서 획득한 샘플들을 저장하고, 재학습하여 샘플 효율성을 향상한 리플레이 메모리(Replay Memory)를 사용한다[7]. 또한, 목표 신경망과 갱신 신경망을 분리하여 신경망 가중치의 갱신을 안정적으로 수행한다. 그러나, DQN은 차원이 낮은 이산(Discrete) 상태를 가진 작업에만 적용할 수 있다.

정책기반 알고리즘은 시스템 상태를 입력으로 신경망을 사용하여 에이전트의 행동을 근사한다. REINFORCE는 대표적인 정책기반 알고리즘으로 몬테카를로 기법을 통해 샘플링한 반환 값을 활용하여 정책신경망을 갱신한다[8]. 샘플 편향은 적지만 한 개의 에피소드(Episode)에서 획득한 반환 값을 사용하기 때문에 샘플의 분산이 큰 특징이 있다.

액터-크리틱 기반 알고리즘은 특정 시스템 상태에서 에이전트의 행동을 근사하는 액터 신경망과 상태의 가치함수 Q를 근사하는 크리틱 신경망을 함께 사용한다. Deep Deterministic Policy Gradient (DDPG)는 결정론적 정책을 사용하여 높은 차원의 연속한 행동공간의 정책 학습이 가능하다[9]. Soft Actor-Critic(SAC)는 DDPG를 확장한 Soft-Q-Learning 기반 알고리즘으로 연속한 행동공간에서의 효율적인 탐색을 수행하고, 지역 최적점에 수렴하는 것을 회피하기 위해 보상함수에 정책의 엔트로피를 추가한다[10]. 본 논문에서는 자율주행 에이전트 구성 시 효율적인 학습 및 탐색 성능을 보이는 SAC 알고리즘을 사용하였다.

2. 종단간 강화학습 자율주행

자율주행을 위한 종단간 강화학습은 차량 센서데이터를 입력으로 모듈화된 파이프라인 없이 직접 주행을 제어한다[1-5]. 결과적으로 시스템의 복잡도가 낮으며 모듈 간 의존성 제약이 적다. 그러나 문제 해결과 에이전트 행동 결정의 해석이 어려운 문제점을 가지고 있다[3-5].

자율주행을 위한 종단간 강화학습은 주로 이미지화된 다양한 센서데이터들을 입력으로 시스템의 상태를 검출하고 이를 입력으로 강화학습 에이전트를 학습한다[1]. 입력 센서의 종류에 따라 차량 센

서데이터만 입력으로 사용하는 방식[2]과 센서데이터와 주행 경로를 입력으로 하는 방식[3]으로 구분할 수 있다. 또한, 강화학습의 연산 복잡도를 줄이기 위해 높은 차원의 센서데이터를 낮은 차원의 잠재 상태(Latent State) 공간으로 변환하는 작업이 필요하다. 이를 위해 변형 자동 인코더(Variational Auto-Encoder: VAE) 등을 사용한다[12]. 본 논문에서는 센서데이터와 주행 경로를 입력으로 변형 자동 인코더로 잠재 상태를 도출하는 논문[3]의 Latent SAC(LSAC)를 기반으로 강화학습 에이전트를 학습한다.

3. 종방향 강화학습 보상함수

강화학습에서 에이전트는 특정 시스템 상태에서 다양한 행동들을 시도하고 그에 대한 보상함수를 경험하면서 점진적으로 보상함수를 최대화하는 방향의 정책을 학습한다. 따라서, 에이전트가 원하는 동작을 수행하게 하기 위해서는 적절한 보상함수의 설계가 필요하다. 특히, 종방향과 횡방향의 다양한 주행 성능, 교차로 신호등 교통 법규의 준수, 교통사고 등의 주행 안전, 그리고 다양한 주행 안락성 지표를 동시에 만족하는 스칼라 보상함수의 설계는 매우 도전적인 주제이다. 본 논문에서는 차량의 종방향 주행성능을 향상하는 새로운 보상함수 설계를 목표로 한다.

자율주행을 위한 기존 강화학습 연구들에서 종방향 보상함수는 특정 운전조작에 대한 인센티브 또는 페널티를 부과하는 형태로 이루어진다. 대부분의 관련 연구들은 교통 효율성 제고를 위해 차량의 속력에 비례하는 인센티브를 주로 제공한다[1-5]. 그러나, 강화학습의 감가율이 1에 근사할 때, 각 시점의 속력보상을 더하면 거리가 되기 때문에 에피소드의 시간이 충분하다면 에이전트가 효율적으로 주행할 유인책이 되지 못한다. 한편, 논문[13]과 [14]는 각각 안전거리 미준수와 빈번한 차선변경 시 에이전트에게 페널티를 부과하는 보상함수를 사용한다. 그러나, 주행 효율성과 안전성을 동시에 고려하여 종방향 보상함수를 체계적으로 설계하고, 성능을 검증한 연구는 지금까지 알려진 바가 없다.

III. 종단간 자율주행 강화학습 에이전트

본 논문에서 차량 좌표축은 차량의 무게 중심을

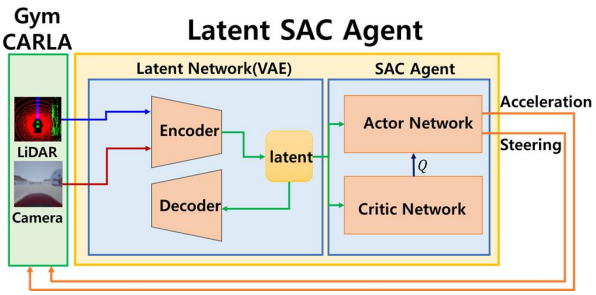

Fig. 1. End-to-end LSAC architecture [3].
그림 1. 종단간 LSAC 신경망의 구조 [3]

원점으로 진행 방향을 X축, 오른쪽 방향을 Y축, 위쪽 방향을 Z축으로 정의한다. 모든 차량은 동일한 길이(l_v)와 폭(w_v)을 가지며, 차량의 동역학적 제한으로 인해 가속도 범위는 $[a_{\min}, a_{\max}]$로 제한한다.

1. Latent SAC 구조

본 논문의 강화학습 에이전트 학습을 위해 CARLA 시뮬레이터를 활용한 LSAC 알고리즘의 강화학습 구조를 사용한다[3]. 그림 1에서 변형 자동 인코더는 CARLA Gym에서 생성된 차량 센서데이터들을 낮은 차원으로 변환하여 잠재 상태를 생성한다. 강화학습 에이전트는 요약된 상태 표현으로 효율적인 학습을 수행할 수 있으며 이를 통해 최적의 정규화된 차량의 조향각과 가속 및 제동 페달 압력을 출력한다.

2. 차량 센서데이터

일반적으로 자율주행차는 주행환경 검출을 위한 카메라, 레이다, 라이다, 초음파 센서데이터들과 위치 감지를 위한 RTK-GPS 수신기, 관성항법장치(Inertial Navigation System: INS) 데이터, 그리고 차량의 CAN(Controller Area Network)으로부터 수신하는 조향각(Steering Angle) 센서, 휠 속력 센서(Wheel Speed Sensor) 데이터들의 종합적인 판단을 통해 주변 차량들과 충돌하지 않는 안전한 경로를 따라 주행할 수 있도록 차량의 조향과 가감속을 제어한다.

차량의 주행 경로가 주어졌을 때 강화학습 입력으로 주어지는 차량 센서들의 종류와 사양은 다음과 같다. 라이다 센서는 감지된 물체의 위치를 3차원 점군(Point Cloud) 데이터 형태로 생성한다. 라이다의 채널 수와 반복률을 각각 C^L과 f^L로, 최대 검출거리를 $D^L_{\max}$로, 수직방향 최대 및 최소 고도각

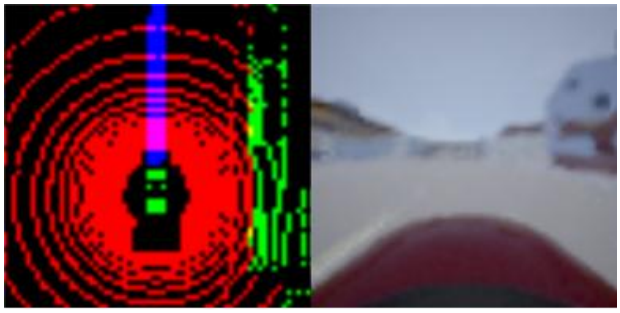

Fig. 2. Input sensor data.
그림 2. 입력 센서데이터

을 각각 $\theta^L_{\min}$과 $\theta^L_{\max}$로 나타낸다. 3차원 라이다의 점군 데이터는 XY 평면에 투영한 조감도(Bird's-Eye View: BEV)로 변환한다. 그림 2 왼편의 라이다 점군 데이터 이미지에서 높이에 따라 도로 노면은 적색으로, 주변 환경 및 이동체는 녹색으로, 차량의 주행 궤적은 파란색으로 표시한다. 중앙의 사진은 전방카메라 이미지를 나타낸다. 그림 2 오른편의 전방 카메라는 수평과 수직 화각이 각각 θ^F_H와 θ^F_V인 공간에 존재하는 물체들의 색상을 감지하여 $P^F_H \times P^F_V$의 해상도를 가지는 이미지를 초당 f^F개 생성한다.

3. 종방향 보상함수

본 절에서는 강화학습 에이전트 차량의 주행 효율성과 안전성을 동시에 고려하면서 전방거리를 적절하게 유지하기 위한 보상함수 설계 방법을 제시한다.

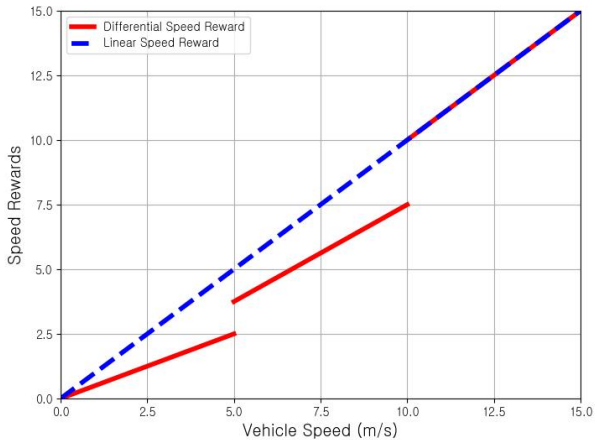

Fig. 3. Linear and differential speed reward vs. speed
그림 3. 속력에 따른 선형속력 보상과 차등속력 보상

그림 3의 청색선으로 표시한 속력 보상은 감가율이 1에 근사하고 에피소드 시간이 충분히 크면 에이전트 차량이 차간거리를 유지하도록 유인하기 위한 대책이 되지 못한다. 결과적으로 자유 흐름(Free Flow) 상황에서도 에이전트 차량은 도로의 제한 속력보다 현저히 낮은 속력으로 주행한다. 도로의 속력 제한이 $v_{\max}$일 때, 차등속력 보상함수는 에이전트 차량의 속력을 $K(=3)$개의 세부 구간으로 분할하고, 각 구간별 속력보상함수 기울기가 점진적으로 증가하도록 ($(s_k < s_{k+1})$) 속력보상을 차등화한다.

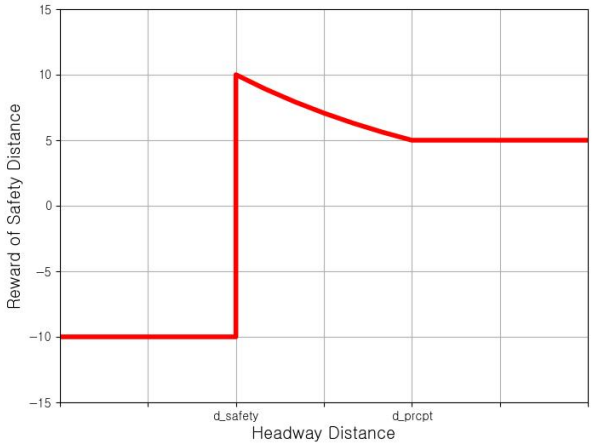

Fig. 4. Safety distance reward vs. headway distance.
그림 4. 차간거리에 따른 안전거리 보상함수

$$R_{speed} = s_k v_{lon}, \text{ for } \frac{(k-1)}{K} v_{\max} \le v_{lon} < \frac{k}{K} v_{\max} \quad (1)$$

그림 3에서 적색선들로 나타낸 차등속력 보상 체계에서 에이전트는 동일한 거리를 주행하더라도 높은 속력으로 주행해야 더 큰 보상을 획득할 수 있으므로 주행 효율성을 향상하기 위한 적절한 유인책으로 볼 수 있다.

한편, 논문[13]에서는 안전거리 미준수 시 에이전트에게 페널티를 부과하여 안전거리 규칙을 따르도록 보상함수를 설계하였다. 그러나, 안전거리 조건을 만족했을 때 주행 효율성을 위해 전방거리를 유지하는 유인책이 부재하다.

자율주행차의 주행 안전성 보장을 위해서는 차량의 최대감지거리 d_{prcpt}가 차량 속력이 $v_{\max}$일 때 안전거리보다 크거나 같아야 한다.

그림 2는 본 논문에서 제시하는 안전거리 보상함수의 예를 차간거리에 따라 나타낸다. 주행 안전성과 효율성을 함께 고려하여 설계한 안전거리 보상함수를 식 (2)에 제시한다.

$$R_{dist}=\begin{cases}-R_{safety}, & d<d_{safety}\\ R_{safety}\exp[-a(d-d_{safety})], & d_{safety}<d<d_{prcpt}\\ 0.5R_{safety}, & d>d_{prcpt}\end{cases} \quad (2)$$

Table 1. CARLA simulation parameters.
표 1. CARLA 시뮬레이션 파라미터

구분	표기	값	단위
Reward Function	$v_{\max}$	15	m/s
	K	3	-
	(s_1, s_2, s_3)	(0.5, 0.75, 1)	-
	R_{safety}	10	-
Vehicle Spec	l_v	4.91	m
	w_v	2.11	m
	$a_{\max}$	3.56	m/s^2
	$a_{\min}$	-8.51	m/s^2
Front Camera Spec	(P_H^F, P_V^F)	(64, 64)	-
	(θ_H^F, θ_V^F)	(110, 40)	°
	f^F	50	FPS
Lidar Spec	C^L	32	-
	$D_{\max}^L$	50	m
	f^L	10	Hz
	$\theta_{\max}^L$	10	°
	$\theta_{\min}^L$	-30	°

만약, 주행 중 에이전트 차량의 차간거리 d가 종방향 속력 v_{lon}에 대한 안전거리 d_{safety}보다 적으면 교통사고를 유발할 수 있기 때문에 $-R_{safety}$를 보상함수로 부여한다. 최대인지거리 내에서 안전거리를 준수하는 에이전트 차량은 주행 효율성 향상을 위해 차간거리가 안전거리에 가까울수록 더 큰 보상을 받도록 설정한다. 경계 조건인 최대인지거리에서 안전거리 보상은 pR_{safety}가 되도록 수식 (2)의 매개변수 a를 설정한다. 마지막으로 인지거리를 초과하면 안전거리 보상을 연속인 상수함수로 정의한다.

Ⅳ. CARLA 시뮬레이션 성능평가

본 장에서는 CARLA 시뮬레이션을 통해 강화학습 에이전트 차량의 종방향 주행 성능을 평가한다. 논문[3]의 LSAC 에이전트의 종방향 보상함수는 종방향 보상함수를 속력으로 나타낸다. 성능 비교를 위해 본 논문에서 제시하는 보상함수를 'Proposed Reward'로, 논문[3]의 보상함수를 'Reward[3]'으로 표기한다.

1. CARLA 시뮬레이션 환경

그림 5는 강화학습 에이전트의 종방향 주행 학습을 위해 생성한 180m 길이의 직진도로 주행환경을 나타낸다. 표 1에서 도로의 제한 속력은 $v_{\max}=15$ m/s며, 강화학습 에이전트의 초기 속력은 10~15 m/s 구간에서 Uniform 분포에 따라 임의로 생성한다. 강화학습 에이전트는 차량이 차선 이탈, 종료지점 도착, 또는 최대주행시간($T_{\max}=50$초)이 경과 시 에피소드를 종료한다.

Fig. 5. Longitudinal driving environment.
그림 5. 종방향 주행환경

차간거리의 학습을 위해 에이전트 차량과 동일한 차선의 전방 20 m에 차량을 한 대 배치한다. 전방 차량의 속력은 10~15 m/s 범위에서 임의의 시간 동안 등속 운동과 가/감속 운동을 교대로 반복한다. 돌발상황에 대처하기 위해 전방 차량은 가/감속 주행 구간에서 확률 $P_{ES}=0.1$로 급정지했다가 다시 출발하도록 주행 속력을 생성한다.

2. 주행 성능 평가

본 절에서는 동일한 차선에서 전방 차량이 속력을 변경해가면서 주행할 때 에이전트 차량의 주행 안전성과 효율성을 평가하기 위해 종방향 속력 프로파일, 안전거리 마진, 전방 차량과의 충돌 확률을 조사한다. 종방향 속력 프로파일은 전체 종방향 거리를 1m 단위로 분할한 후, 에이전트 차량이 해당 위치에 있을 때 속력 측정값들의 평균으로 나타낸다. 안전거리 마진은 전방 차량과의 차간거리에서

안전거리를 차감한 값으로 정의하고, 전방 차량 충돌 확률은 에이전트 차량의 전체 행동 중에서 전방 차량과 충돌을 유발한 행동의 비율로 정의한다.

그림 6은 종방향 차량 위치에 따른 에이전트 차량의 속력 프로파일을 나타낸다. 논문[3]의 에이전트 차량은 보상함수에서 속력의 제곱에 비례하는 패널티로 인해 차량 속력의 평균이 6.37 m/s로 제한된다. 반면, 동일한 상황에서 제시하는 보상함수를 기반으로 학습한 에이전트 차량은 속력 구간별로 차등화된 가중치를 활용하여 보상함수를 생성하기 때문에 10.75 m/s의 고속 주행이 가능하다. 하지만, 제안하는 보상함수를 사용하면 차간거리의 유지를 위해 빈번한 가감속을 수행하는 문제점이 있으며, 이를 해결하기 위한 추가 연구가 필요하다.

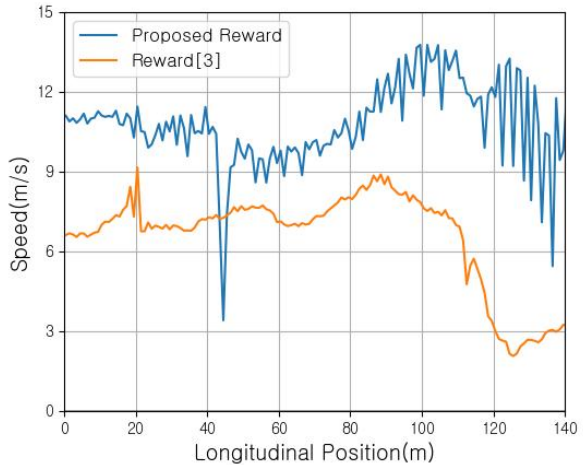

Fig. 6. Longitudinal speed profile.
그림 6. 종방향 속력 프로파일

그림 7은 에이전트 차량의 안전거리 마진을 나타낸다. 논문[3]의 에이전트 차량은 주행 속력의 한계로 인해 전방거리가 크게 증가하더라도 안전거리를 줄이지 못한다. 결과적으로 평균 안전거리 마진이 37.52 m로 주행 효율성이 크게 저하되는 문제점이 발생한다. 반면, 제시하는 보상함수로 학습한 에이전트 차량은 평균 안전거리 마진이 -1.78 m로 돌발상황에서 전방 차량이 최대로 제동하면 충돌의 위험이 배제하지 못하지만, 일반적인 주행환경에서 전방 차량을 잘 추종하며 이동함을 확인할 수 있다. 다시 말해, 새로운 보상함수로 학습한 에이전트 차량은 운전자들이 상황에 따라 일부 안전거리를 위반하더라도 적응적으로 차간거리를 조정하는 방식과 매우 유사하게 차량의 속력을 제어함을 알 수 있다.

마지막으로 두 에이전트 차량의 전방 차량과 충돌 확률은 예상과 다른 결과를 나타낸다. 충분한 안전거리 마진을 가지는 논문[3] 에이전트 차량의 전방 차량 충돌 확률이 약 12 %인 반면 제시하는 방식의 에이전트 차량은 모든 상황에서 전방 차량과 충돌하지 않았다. 이 결과는 두 가지 측면에서 추론할 수 있다. 첫째, 제시하는 방식의 에이전트 차량이 15.08 m의 안전거리 마진 표준편차를 가지는 반면 논문[3]의 에이전트 차량은 두 배가 넘는 31.88 m 안전거리 마진 표준편차를 가지기 때문에 안전거리를 위반하더라도 충돌사고를 회피하면서 차간거리를 유지할 수 있다. 둘째, 제안하는 방식의 에이전트 차량은 안전거리 마진이 0보다 적어지면 높은 압력으로 제동 페달을 밟는 것으로 추론된다. 이는 그림 6에서 종방향 주행속력의 급격한 변동을 일으키는 주요 원인이 되지만 일정한 차간거리를 유지하면서 주행 안전성을 지킬 수 있는 제어 방법이다.

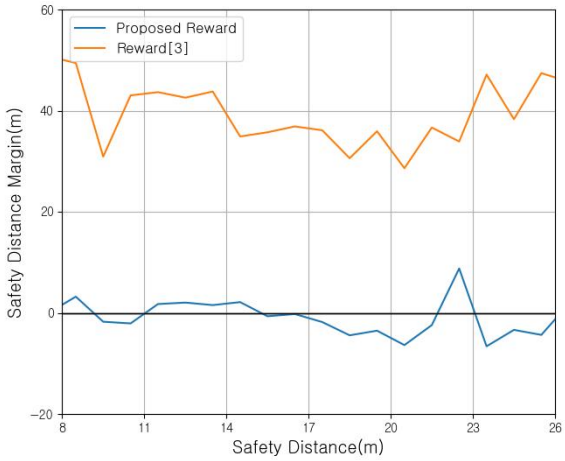

Fig. 7. Safety distance margin.
그림 7. 안전거리 마진

Ⅴ. 결론

본 논문은 자율주행을 위한 잠재 SAC 기반의 심층강화학습 에이전트의 종방향 주행 안전성과 효율성을 향상하는 강화학습 보상함수 체계를 제시한다. CARLA 시뮬레이션을 통해 제시하는 방식의 에이전트 차량이 높은 수준의 종방향 주행특성을 달성할 수 있음을 보였다. 향후, 본 논문의 보상함수 체계를 확장하여 종방향과 횡방향에서 높은 주

행 성능을 달성하고 신호등 규칙을 준수하는 강화 학습 에이전트를 개발할 예정이다.

References

[1] Z. Zhu and H. Zhao, "A survey of deep RL and IL for autonomous driving policy learning," *arXiv preprint,* arXiv:2101.01993, 2021.

[2] H. Abdou et al, "End-to-end deep conditional imitation learning for autonomous driving," *Proc. of IEEE ICM'19*, pp.346-334, 2019.

[3] M. Bansal, K. Alex, and O. Abhijit, "Chauffeurnet: Learning to drive by imitating the best and synthesizing the wors," *arXiv preprint arXiv*: 1812.03079, 2018.

[4] W. Zeng et al. "End-to-end interpretable neural motion planner," *Proc. of the IEEE CVPR'19*, 2019.

[5] J. Chen, E. L. Shengbo, and T. Masayoshi, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Trans on Intelli. Transpt. Syst.*, 2021.

[6] A. Dosovitskiy et al. "CARLA: An open urban driving simulator," *Conf. on Robot Learning.* 2017.

[7] V. Mnih et al. "Human-level control through deep reinforcement learning," *Nature*, vol.518, no.7540 pp.529-533, 2015.

[8] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol.8, no.3, pp.229-256, 1992. DOI: 10.1007/BF00992696

[9] T. P. Lillicrap et al. "Continuous control with deep reinforcement learning," arXiv preprint, arXiv: 1349.02971, 2015.

[10] T. Haarnoja et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *Intern. Conf. on Machine Learning*, 2018.

[12] D. P. Kingma, and W. Max, "Auto-encoding variational bayes," *arXiv preprint* arXiv:1312.6114, 2013.

[13] D. Zhao, Z. Xia, and Q. Zhang, "Model-free optimal control based intelligent cruise control with hardware-in-the-loop demonstration," *IEEE Comput. Intelli. Mag.*, vol.12, no.2, pp.56-69, 2017.

[14] C. Desjardins and B. Chaib-Draa, "Cooperative adaptive cruise control: A reinforcement learning approach," *IEEE Trans. on intelli. transpt. syst.*, vol.12, no.4, pp.1248-1260, 2011.

BIOGRAPHY

Sung-Bean Jo (Member)

2020 : BS degree in Electrical Engineering, Pusan National University.
2020~present : MS degree in Electronic and Electrical Engineering, Pusan National University.

Han-You Jeong (Member)

1998 : BS degree in Electrical Engineering, Seoul National University.
2000 : MS degree in Electrical and Computer Engineering, Seoul National University.
2006 : PhD degree in Electrical and Computer Engineering, Seoul National University.
2014~present : Professor, Pusan National University