# PPO기반 자율주행차량의 효율적이고 안전한 차선 변경 정책 연구

이동수, 권민혜
숭실 대학교 브레인 및 기계 지능 연구실
movementwater@soongsil.ac.kr, minhae@ssu.ac.kr

# PPO-based Efficient and Safe Lane Change Strategy for Autonomous Vehicles

Dongsu Lee, Minhae Kwon
Brain and Machine Intelligence Laboratory, Soongsil University

요 약

자율주행기술의 발전과 함께 상용화가 이루어지고 있는 과정에서 자율주행차량의 안전성과 신속한 주행 능력이 요구되고 있다. 특히 다수의 차선이 존재하는 도로 상황의 경우, 차선 변경을 위한 의사결정이 필수적인데 이는 현재 차선의 전 후방 차량의 상황 뿐 아니라 이동하고자 하는 차선에 위치한 차량들의 상황 또한 고려해야 하기에 그 의사결정의 난이도가 매우 높다. 본 논문에서는 효율적이고 안전한 차선 변경에 대한 마르코프 의사결정 모델을 제안한다. 제안한 모델을 통해 심층 강화학습 기반 알고리즘인 PPO를 이용하여 자율주행차량을 학습한다. 학습된 차량은 제어이론 기반의 차선 변경 모델과의 성능 비교를 통해 보다 효율적이고 안전하게 차선 변경을 수행하는 모습을 확인하였다.

## Ⅰ. 서 론

자율주행기술 발전의 가속화가 진행되며 심층 강화 학습 기반의 자율주행 차량 제어 기술이 활발히 연구되고 있다. 자율주행차량이 상용화되기 위해서는 현실적인 도로 상황을 대비하는 기술의 필요성이 요구된다. 본 연구에서는 다수의 차선이 존재하는 도로 환경에서 자율주행차량이 차선 변경을 수행할 수 있도록 학습하고자 한다. 차선 변경은 단순한 가속도 제어 연구와는 달리, 다수의 차선에 존재하는 전 후방 차량을 모두 고려해야 하기에 그 난이도가 높은 의사결정과정이다. 차선 변경의 수행이 잘못될 경우 도로의 정체 유발 및 안전사고를 일으킬 수 있기에 자율주행차량에서는 필수적으로 연구되어야 하는 기능이다. 본 논문에서는 신속한 주행 및 안정성을 위한 마르코프 의사결정 과정(Markov Decision Process; MDP)을 모델링하여, 심층 강화학습 알고리즘인 Proximal Policy Optimization(PPO)[1]를 통해 자율주행차량을 학습시켜 효과를 확인하고자 한다.

## Ⅱ. 차선 변경 학습을 위한 심층강화학습 기반 모델 설정

강화학습은 학습의 주체인 개체가 환경과의 상호작용을 통해 학습을 하는 기계학습의 방법 중 하나이다. 강화학습 문제는 MDP를 따른다. MDP는 개체가 수행하는 의사 결정 과정을 확률적으로 모델링하는 방법으로 일련의 튜플 $< S, O, A, R, \gamma >$로 정의할 수 있다. 상태 공간(state space) $S$는 개체가 상호작용하는 환경의 시간 $t$에서의 상태(state) $s_t$의 집합이다. 관측 공간(observation space) $O$는 개체가 환경을 관측 정보(observation) $o_t$의 집합이다. 이때 개체가 관측 가능한 상태 정보의 집합이 상태 공간과 동일한 경우 완전 관측(full observation)이라고 하며, 일부로 한정되는 경우를 부분 관측(partial observation)이라고 한다. 행동 공간(action space) $A$는 개체가 취할 수 있는 모든 행동(action) $a_t$의 집합이다. 보상함수 $R(s_t, a_t, s_{t+1})$(이하 $R_{t+1}$로 표기)은 상태 $s_t$에서 행동 $a_t$를 취할 때 변한 상태 $s_{t+1}$에 대해 환경이 개체에게 주는 보상을 의미한다. 개체는 특정 상태 $s_t$에서 보상 $R_{t+1}$가 최대가 되는 행동 $a_t$를 취하는 방향으로 학습한다. 마지막으로 $\gamma$는 시간에 따른 감가율(discount factor)을 의미한다.

### Ⅱ.1. 도로 환경 설정

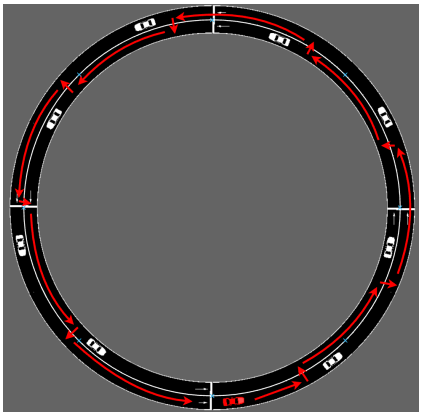

그림 1. 이차선 원형도로 구조 및 자율주행차량의 이동경로

본 연구에서는 개체가 차선 변경 학습을 충분히 수행할 수 있는 그림 1과 같은 2차선 원형도로를 다룬다. 도로 내 비 자율주행차량들은 느리게 등속 주행을 하여 자율주행 차량이 차선 변경을 수행하여야만 목표 속도에 도달할 수 있는 환경을 조성하였다.

도로 내 차량의 집합 $E = [e_1, e_2, \cdots, e_N]$은 비 자율주행차량의 집합 $E_{non} = [e_1, e_2 \cdots, e_{N-1}]$와 자율주행차량의 집합 $E_{rl} = [e_N]$으로 구성한다. 도로에 배치된 전체 차량의 수 $|E| = N$ 대이다. 차선 번호 $k$는 가장 오른쪽 차선이 0번 차선이며 왼쪽으로 갈수록 차선의 번호는 증가한다.

### Ⅱ.2. 차선 변경을 위한 마르코프 의사결정 과정

자율주행차량 $e_N$은 도로 전체의 상태 정보 $s_t$가 아닌 근접 도로 상태에 대한 부분 관측만이 가능하다. $e_N$의 관측 정보 $o_t \in O$은 다음과 같은 14차

원으로 정의한다.($o_t \in \mathbb{R}^{14\times 1}$)

$$o_t = [v_{t,l0}, v_{t,l1}, v_{t,f0}, v_{t,f1}, v_{t,N}, p_{t,l0}, p_{t,l1}, p_{t,f0}, p_{t,f1}, k_{t,lo}, k_{t,l1}, k_{t,f0}, k_{t,f1}, k_{t,N}]^T$$

여기서 $[v_{t,l0}, v_{t,l1}, v_{t,f0}, v_{t,f1}, v_{t,N}]$는 각각 0번 차선 선두차량, 1번 차선 선두차량, 0번 차선 후방차량, 1번 차선 후방차량 그리고 자율주행차량의 속도를 의미한다. $[p_{t,l0}, p_{t,l1}, p_{t,f0}, p_{t,f1}]$는 각각 0번 차선 선두차량, 1번 차선 선두차량, 0번 차선 후방차량, 1번 차선 후방차량과 자율주행차량 사이의 상대 거리를 의미한다. 마지막으로 $[k_{t,lo}, k_{t,l1}, k_{t,f0}, k_{t,f1}, k_{t,N}]$는 0번 차선 선두차량, 1번 차선 선두차량, 0번 차선 후방차량, 1번 차선 후방차량 그리고 자율주행차량의 차선을 의미한다.

개체가 취할 수 있는 행동 $a = \{acc, lc\}$로 나타낸다. $acc$는 자율주행차량의 가속도를 의미하며, $acc \in [-1,1]$의 연속적인 범위를 갖는다. $lc$는 자율주행차량의 차선 변경 방향을 의미하며, $lc \in \{-1,0,1\}$와 같은 이산적인 값을 갖는다. 이때 0은 차선을 유지하는 경우, -1은 우측 차선으로의 차선 변경, 1은 좌측 차선으로의 차선 변경을 의미한다.

차선 변경을 통해 개체가 효율적인 주행을 할 수 있도록 하며 동시에 주변 차량의 주행을 방해하지 않기 위한 보상 함수는 다음과 같다.

$$R_{t+1} = \eta_1\left(1 - \left|\frac{v_{t+1,N} - v^*}{v^*}\right|\right) - \eta_2\left(\min\left[0, 1 - \left(\frac{s^*}{p_{t+1,f}}\right)\right]\right) \quad (1)$$

$\left(1 - \left|\frac{v_{t+1,N} - v^*}{v^*}\right|\right)$는 보상 항으로 자율주행차량이 목표 속도 $v^*$에 가깝게 주행할 수 있도록 한다. 만약 $v_{t+1,N}$이 목표 속도 $v^*$와 동일하다면 최대 보상 1이 주어지며, $v^*$에서 증가하거나 감소하는 경우 그보다 낮은 보상이 주어진다. $\left(\min\left[0, 1 - \left(\frac{s^*}{p_{t+1,f}}\right)\right]\right)$는 자율주행차량이 차선 변경 했을 때 후방 차량의 안전 범위를 침범하는 것에 대한 처벌 항이다. $p_{t+1,f}$는 시간 $t+1$에서 후방차량과 자율주행차량 사이의 상대 거리를 의미한다. $s^*$는 안전거리이며 이는 환경 설정 및 사용자에 의해 조절될 수 있다. 본 연구에서 비 자율주행차량의 안전거리는 IDM 컨트롤러[2]에 의해 조절되기 때문에 $s^* = s_0 + \max\left(0, v_{t+1,f}(t^* + \frac{v_{t+1,f} - v_{t+1,N}}{2\sqrt{acc_{\max} acc_{\min}}})\right)$와 같이 설정하였다. $s_0$는 차량 간 최소 허용 거리이며, $t^*$은 time headway로 선두차량과 후방차량이 동일한 위치에 도달하는데 필요한 최소 허용 시간이다.

### Ⅱ.3. FLOW 시뮬레이터 설정

본 연구의 성능 평가를 위해 도로 교통 시뮬레이터에 대한 심층 강화학습 프레임워크 FLOW[3]를 사용하였다. 도로의 구성은 260m의 2차선 원형 도로(그림 1)이며 차량의 수 $|E| = 9$대 이다. 여기서 자율주행차량의 수 $|E_{rl}| = 1$대 이며 비 자율주행차량의 수 $|E_{non}| = 8$대 이다. 비 자율주행차량은 모두 IDM 컨트롤러를 사용하며 주행 속도는 $1m/s$로 고정하였다. 최소 허용 거리 $s_0 = 2m$, time headway $t^* = 1ts$, 목표 속도 $v^* = 3m/s$로 설정하였다. 본 시뮬레이션에서 수식 (1)의 $\eta_1$, $\eta_2$는 10, 1로 설정하였으며 1 time step $ts = 0.1s$로 정의하였다.

## Ⅲ. 차선 변경 학습 모델 평가

심층 강화학습 알고리즘 PPO로 학습한 차량의 성능 평가를 위해 제어이론 기반의 LC2013[4] 차선 변경 모델을 적용한 경우와 성능 비교를 진행하였다. 그림 2를 통해 두 차량 모두 목표 속도인 $3m/s$를 유지하려는 모습을 확인할 수 있다. 표 1을 통해 단일 에피소드에서 차선 변경을 수행하는 단일 차량의 평균 속도와 속도의 분산을 자세히 확인할 수 있다. LC2013 모델을 사용한 경우 차선 변경 결정을 바로 내리지 못하고 비 자율주행 선두차량 뒤에서 선두차량의 속도에 맞춰 운행하며 시간을 지체하는 것을 확인할 수 있다. 이는 그림 2의 time step 1700 및 2500 부근에서 비 자율주행차량들의 속도인 $1m/s$로 유지하는 모습을 통해 단적으로 확인 가능하다. 반면 PPO를 사용하는 경우에는 전방 차량이 길을 막고 있을 때 의미 없이 기다리지 않고 차선을 변경하기 때문에 일정한 속도를 유지하며 주행하는 모습을 확인하였다. 결과적으로 제어 이론 기반의 차선 변경 모델인 LC2013을 사용한 경우에 비해 심층강화학습 기반 모델인 PPO를 사용하여 학습한 경우 더욱 자연스러운 차선 변경을 수행하는 모습을 확인하였다.

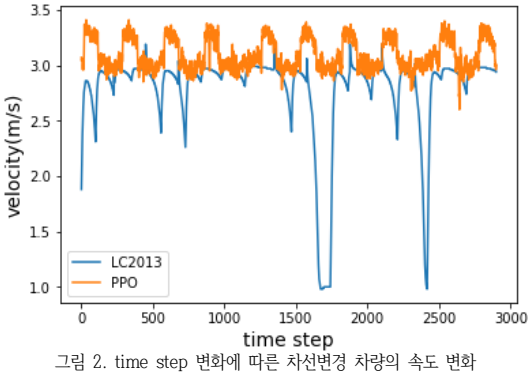

그림 2. time step 변화에 따른 차선변경 차량의 속도 변화

표 1. 단일 에피소드에서의 차선변경 수행 차량의 평균 속도 및 속도의 분

| Algorithm | Avg. velocity | Variance |
|---|---|---|
| PPO | $3.09m/s$ | 0.019 |
| LC2013 | $2.78m/s$ | 0.168 |

## Ⅳ. 결론

본 논문에서 우리는 자율주행차량의 차선변경 학습을 위한 MDP 모델을 제안하였다. 심층 강화학습 알고리즘인 PPO를 통해 학습한 자율주행차량은 전통적인 제어이론 기반 차선 변경 모델을 적용한 비 자율주행차량과 비교하였을 때 더욱 높은 성능을 보여주었다. 자율주행차량의 평균 속력은 비 자율주행차량 약 10% 상승 하였고 목표 속도에 가까운 주행 능력을 보였다.

### Acknowledgement

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터 지원사업(IITP- 2021-2020-0-01602)과 방송통신산업기술개발사업(IITP-2021-0-00739), 그리고 한국 연구재단(NRF- 2020R1F1A1069182)의 지원을 받아 수행된 연구임.

### 참 고 문 헌

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

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

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

[4] J. Erdmann, "SUMO's Lane-Changing Model," Springer, 2015.