탈중앙화 심층강화학습 기반 자율주행 지역경로 선택 및 속도 제어 기술

강민수, 이일규, 조윤식, 신오순
숭실대학교 전자정보공학부 IT 융합전공

{dorae97, xignos3108, yunsik96}@soongsil.ac.kr, osshin@ssu.ac.kr

Route Selection and Speed Control of Autonomous Vehicles Based on Decentralized Deep Reinforcement Learning

Minsoo Kang, Ilkyu Yi, Yunsik Cho, Oh-Soon Shin
Soongsil University

요 약

자율주행은 운전자의 개입없이 주변환경을 인식하고, 주행상황을 판단하여 자동차를 제어하는 높은 정밀도를 가진 시스템으로 발전되어왔다. 하지만 고전적인 지도학습 방법으로는 복잡하고 유동적인 도로환경에 맞추어 최적의 행동을 도출하는데 한계가 있다. 본 논문에서는 심층강화학습 알고리즘인 Deep Q Network 를 이용한 지역경로 선택 및 Deep Deterministic Policy Gradient 를 이용한 속도 제어를 수행하는 자율주행 시스템을 제안하고 모의실험을 통해 성능을 평가한다.

Ⅰ. 서 론

자율주행 실현을 위해서는 정확한 판단과 명령이 중요하므로 최근 강화학습을 도입하여 자율주행의 성능과 안정성을 개선하기 위한 다양한 연구들이 진행되었다[1]. 하지만 속도와 주행 계획의 동시성이 중요한 자율주행 시스템의 특성상 가속도 제어를 통한 자동차의 속도 조절, 교차로에서의 방향 결정, 전역 경로 설정 등의 순차적, 독립적인 접근방식은 현실에 적응하기 어려웠다. 또한, 5G 의 확산으로 서버를 통해 다수의 Agent 를 동시에 운영하기 위한 시도를 반영하여 Single-agent 강화학습(Reinforcement Learning)을 각 Agent 마다 부여하였다. 그러나 이러한 접근은 각 Agent 의 정책 차이로 인해 유동흐름에 부정적인 영향을 미칠 수 있다. 반면, Multi-agent 강화학습의 경우 협력적인 행동으로 주변 환경에 긍정적인 영향을 미칠 수 있으나, Agent 가 증가함에 따라 행동 및 상태공간이 넓어져 차원의 저주에 빠질 가능성이 있다.

기존의 Single-agent 및 Multi-agent 강화학습의 문제 해결을 위해 본 논문에서는 모든 자동차가 동일한 정책을 갖는 탈중앙화(Decentralized) 강화학습[2] 방식을 통해 문제 해결을 시도하였다. 자동차가 교통 시뮬레이터인 SUMO[3]를 활용하여 교통상황이 존재하는 도로망을 구현[4]하고, 자율주행 자동차에서 강화학습 알고리즘인 Deep Q Network (DQN)[5] 와 Deep Deterministic Policy Gradient (DDPG)[6]를 동시에 활용하여 최적의 경로를 찾는 모델을 제안하고 성능을 평가한다.

Ⅱ. 심층강화학습 기반 자율주행 기술

1. 강화학습 요소 설정

자율주행 시스템 구축에 있어서 교통상황의 복잡성 및 유동성은 알고리즘 구성과 이를 뒷받침하는 연산의 비용 증가를 유발한다. 복잡하고 다양한 상황에서 최적의 판단에 따른 대응이 가능한 강화학습이 자율주행 시스템에 효과적으로 적용될 수 있다. 강화학습은 Markov Decision Process (MDP)에 의거하여 임의의 환경을 포함한 상태(State)에서 학습 주체인 Agent 의 행동(Action)에 보상(Reward)을 부여하여 점진적으로

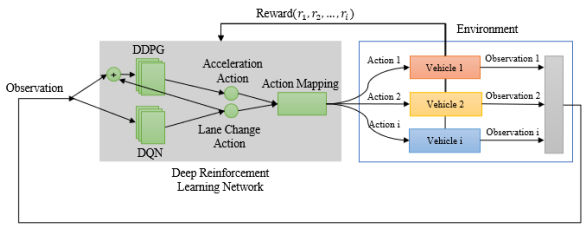


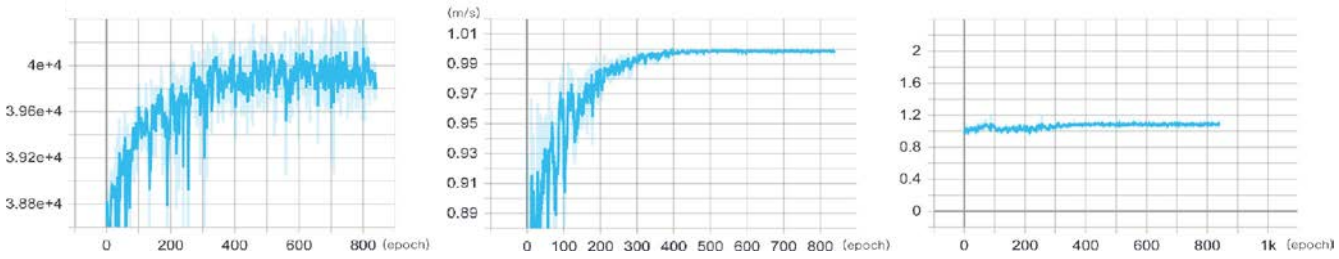
그림 1. 제안하는 심층강화학습 모델.

최적의 결과를 도출하는 학습모델이다. 본 논문에서 고려하는 Agent 인 자율주행 자동차에 대한 강화학습의 각 요소는 다음과 같이 설정하였다.

1) 행동: 자율주행에서 행동은 지역경로 판단과 속도 결정으로 구분될 수 있다. 지역 경로 판단은 목적지로의 조향을 위한 차선변경을 의미하며 우현 차선 변경, 직진, 좌현차선변경을 각각 0, 1, 2 로 구분하였다. 속도 결정은 현실감 있는 사용자 경험을 고려하여 주행 중 가속페달과 브레이크를 밟는 행위를 모사한 현재 속도에서의 가속도로 설정하였다. 감속과 가속의 상황을 고려하여 $[-1, 1]m/s^2$ 구간의 공간을 부여하였다.

2) 상태: Agent 는 환경 내에 위치하여 일부의 상태만 관찰할 수 있는 부분적 관찰가능(Partially Observable) 상황을 가정하였다. Agent 가 관찰하는 상태는 운전자가 관찰하는 환경을 고려하여 총 9 개(현재 속도, 우현차선변경 가능 여부, 좌현차선변경 가능 여부, 선행자동차와의 거리, 후행자동차와의 거리, 현재 차선, 경로내 현재위치, 이동방향, 현재교통신호)로 설정하였다.

3) 보상: Agent 의 행동에 대한 성능은 보상과 손실의 합으로 평가한다. 자율주행의 목적은 적절한 경로로 목적지까지 안전하게 최단시간으로 이동하는데 있으므로, 보상은 각 Time Step 의 속도 합으로 지정하였다. 단, 학습에서 발생하는 차선의 변경은 과도하면 바람직하지 않다는 점에서 -1 의 Penalty 를 부과하였다. 또한 SUMO 환경에서 충돌이나 긴급정지(Emergency Stop) 등의 돌발 행동이 발생할 때, 시뮬레이션의 붕괴 방지를 위해 Agent 를 임의로 이동(Teleport) 시킨다. 이는 현실에서 안전 사고를 유발하므로 해당 도로 최대속력의 절반을 Penalty 로 부여하여 방지할 수 있도록 하였다.



(a) (b) (c)

그림 2. 모의실험 결과. (a) Reward 변화, (b) DDPG 가속도 평균값 변화, (c) DQN 차선변경 시도 행동의 평균값 변화.

2. 심층강화학습 모델

그림 1 과 같이 지역 경로 판단을 위한 DQN 알고리즘과 속도 제어를 위한 DDPG 알고리즘을 이용하여 심층강화학습 모델을 구축하였다. 학습 과정인 State Backpropagation 시 서로 영향을 미치지 않도록 설계하여 독립성을 부여하였으나, DQN 알고리즘의 행동 도출 시 동시에 DDPG 알고리즘이 도출하는 행동에 영향을 미치도록 하여 유기적인 소통이 가능한 구조로 설계하였다.

1) DQN 기반 지역경로 판단: Agent 가 도출하는 행동 중 지역경로 판단은 각 Time Step 에서의 차선변경으로 나타난다. DQN 은 상태의 입력에 의해 가능한 모든 행동의 예측 값으로써 이산행동공간에 존재하는 결과를 도출하는데 용이하다. DQN 은 탐욕적 행동(Greedy Action) 정책을 따라 점수를 예측하고 목적함수를 최적화하여 다음 행동을 선택한다. 목적함수는 Bellman 방정식에 따라 선택되며 목적 네트워크(Target Network)로부터 예측된 Q 값으로 최적화된다. 이는 식 (1)과 같이 표현된다.

$$J = (r_t + \gamma \max Q'(s_{t+1}, a_{t+1}) - Q(s_t, a_t))^2 \quad (1)$$

2) DDPG 기반 속도 제어: 속도결정의 경우 연속적인 가속도 조절을 통해 진행되므로 연속행동공간에서 사용가능한 DDPG 를 통해 이를 해결하고자 했다. DDPG 은 Actor-Critic 모델에 기반한 학습 알고리즘으로 Critic 은 DQN 의 Bellman 방정식을 통해 Update 하였고, Actor 는 Policy Performance 의 기울기 방향(Gradient Ascent)으로 Update 한다. 상태공간에 대한 일반화를 위해 Neural Function Approximator 를 사용하였으며 목적함수의 Update 는 식 (2)와 같다.

$$\nabla_\theta J = E\left[\nabla_a (Q(s,a|\theta^Q))\big|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu} \mu(s|Q^\mu)\big|_{s=s_t}\right] \quad (2)$$

III. 모의실험 결과

모의실험에서 고려한 환경은 신호등이 존재하는 열십자(十) 도로로서 적절한 수준의 교통상황이 존재하도록 한다. 에피소드 시작 후 총 70 개의 Agent 가 50 Time Step 간격으로 추가될 수 있도록 하여 자동차가 안정적으로 주행할 수 있도록 하였다.

모의실험 결과를 그림 2 에 제시하였다. 그림 2(a)의 Reward 변화 그래프를 보면 학습이 진행됨에 따라 Reward 가 점진적으로 개선되어 수렴하는 경향을 나타냄을 확인할 수 있다. 이를 통해 점진적인 성능향상과 안정성의 개선이 있었음을 유추할 수 있다. 과도한 차선변경이나 긴급정지 등의 횟수도 학습 초기 13 회에 대비하여 없어져서 안정적인 주행을 하는 것을 확인할 수 있었다. Agent 의 학습에 따른 행동 경향성 파악을 위해 그림 1(b)을 보면, DDPG 행동인 가속도의 평균값이 $1m/s^2$ 에 수렴하는 것을 확인할 수 있다. 이는 Agent 가 가속을 통해 속도를 최적화하려는 것으로 해석된다. 반면, 그림 1(c)에 제시한 DQN 의 차선변경 시도 행동의 평균값은 예측 값인 1 을 초과하는 수치로 수렴되어 학습결과 일정한 방향으로 치우쳐 차선변경을 진행하는 경향이 있음을 확인하였다. 이는 무작위로 도출된 Agent 의 목적지가 특정 조향으로 치우쳐 나타난 경우와 실제로 주변 자동차의 진행 방향이 좌측에 더 여유공간이 있거나 Agent 생성시에 좌측 자동차의 진행이 적은 경우를 고려할 수 있다. 이에 대해서는 추후 일반 자동차 생성과 Agent 자동차 생성 시기의 관계를 분석하여 연구를 진행할 계획이다.

IV. 결론

본 논문에서는 심층강화학습 알고리즘인 DQN 과 DDPG 모델을 동시에 사용하여 자율주행 성능을 개선하는 연구를 진행하였다. 또한, Multi-agent 에 적용 가능한 구조를 제시하기 위해 탈중앙화 강화학습을 사용하여 현실에 적용이 쉽고 Scalability 를 갖춘 모델을 제시하였다. 향후에는 대규모 강화학습 기반 자율주행 자동차들이 존재하는 다양한 도로 환경에서 동작 가능한 연구를 진행하고자 한다.

Acknowledgment

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019R1A2C1084834).

참 고 문 헌

[1] A. Haydari and Y. Yilmaz, "Deep reinforcement learning for intelligent transportation systems: A survey," *IEEE Trans. Intelligent Transportation Systems*, 2020.

[2] C. Chen *et al.*, "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 4, pp. 3414-3421, 2020.

[3] P. A. Lopez *et al.*, "Microscopic traffic simulation using SUMO," in *Proc. IEEE Inter. Conf. Intelligent Transport Systems (ITSC)*, pp. 2575-2582, 2018.

[4] https://github.com/3neutronstar/sumo_edgecloud

[5] V. Mnih *et al.*, "Playing Atari with deep reinforcement learning," *arXiv Preprint*, arXiv:1312.5602, 2013.

[6] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *arXiv Preprint*, arXiv:1509.02971, 2015.