자율주행 차량의 제한속도 준수를 위한 심층강화학습기반 속도제어방법 연구

최석웅 한상익 권민혜
숭실대학교
{scsc0511,sangikhan}@soongsil.ac.kr minhae@ssu.ac.kr

Deep Reinforcement Learning Based Velocity Controls for Autonomous Vehicles with Speed Limits

Seok-Ung Choi Sang-ik Han Minhae Kwon
Soongsil University

요 약

본 논문에서는 자율 주행 자동차의 규정 속도 준수를 위해 대표적인 정책경사기반 심층 강화 학습 알고리즘인 Proximal Policy Optimization을 이용하여 학습시키고, 그 성능을 평가하였다. 도로 구간 별 규정 속도 변화가 있는 환경에서 자율 주행 차량이 스스로 속도를 제어하는 것을 최종 목표로 하였다. Carla 시뮬레이터를 이용하여 실험한 결과 3 가지의 도로 별 규정속도에 맞춰 자율 주행 자동차가 속도제어를 하는 것을 확인하였다. 특히 규정 속도가 빠르게 변하여 차량의 속도를 규정 속도에 맞출 시간이 짧은 경우에도 이를 잘 준수하여 주행하는 것을 확인하였다.

Ⅰ. 서 론

최근 심층 신경망을 기반기술이 접목되면서 많은 발전을 이루었다. 이로 인해 고속도로 주행 보조, 원격 스마트 주차 보조 등의 기술이 상용화 되어 실생활에서도 사용되고 있다. 하지만 현재 시점에서 자율 주행 기술이 인간의 운전능력을 대체하기에는 자율 주행 차량의 상황에 적합한 운전 능력의 부족 등의 문제가 존재한다. 국토 교통부와 한국 교통 연구원의 연구에 따르면 법정 제한 속도는 도로 여건, 교통량, 사고 유형 등을 종합해서 결정되므로 제한 속도를 준수하는 것은 차량이 처한 상황에 맞게 운전을 하고 있는 지에 대한 척도로서 적절하다고 할 수 있다[1]. 따라서 본 연구는 심층 강화 학습의 Policy Gradient 방법의 하나인 Proximal Policy Optimization(PPO)[2]를 사용하여 자율 주행 차량의 제한 속도를 준수할 수 있게 하는 방법을 제안하고자 한다.

Ⅱ. 법정 제한 속도 준수를 위한 심층 강화학습 방법

개체의 제한 속도 준수라는 목적을 달성하기 위해서 Markov Decision Process (MDP)를 사용하여 문제를 정의하였다. MDP는 환경과 개체 사이의 상호 작용이 연속적으로 이루어지는 상황에 대해 Markov Property를 기반으로 단순화한 확률 기반 의사 결정 모델이다. Markov Property는 현재 상태가 다음 상태를 예측하기 위해 필요한 충분한 정보를 가지고 있는 속성을 의미한다.

MDP는 상태 공간, 행동 공간, 트랜지션 모델, 보상 모델, 감인으로 구성된다. 상태 공간은 개체를 둘러싼 환경에 대하여 개체가 인식할 수 있는 임의의 가능 상태들의 집합이다. 본 논문에서는 자율 주행 차량으로 서의 개체가 전체 상태의 일부분만 관측이 가능하다고 가정하며 개체가 관측 가능한 상태로는 개체의 속도, 제한 속도, 개체의 위치, 차선 중심에 위치하는 Waypoint의 위치, 개체의 방향, 차선의 방향과 개체의 충돌여부를 알려주는 collision flag를 추가하여 개체의 상태정보를 정의하였다. 이 내용은 그림2에 설명을 추가하였다.

개체의 행동은 가속 페달을 밟는 것과 브레이크를 밟는 정도로 나타냈다. 각각은 0에서 1사이의 값으로 연속적인 행동공간을 가진다. 이를 위해 신경망의 출력 층의 활성 함수로 tanh 함수를 사용하여 -1에서 1사이의 값을 출력한 후 0에서

그림 1. Carla 시뮬레이터의 Town01 지도의 모습

1사이의 값일 경우 가속 페달을 밟는 정도를 나타내고 -1에서 0사이의 값일 경우 절댓값을 취한 값을 브레이크를 밟는 정도로 나타냈다. 차량의 steering은 Carla에서 기본적으로 제공되는 autopilot mode에서 차량을 제어하는 방법인 proportional -integral-derivative (PID) [3] controller를 사용하였다.

개체의 보상 모델은 특정 상태에서 개체가 수행한 어떤 행동으로 인해 나온 결과에 따라 결정되는 함수로서 개체가 어떻게 행동하는 것이 좋은지에 대해 알려준다.

$$R = R_s + R_d \quad ---(1)$$

$$R_s = -((\mathrm{S_d} - \mathrm{S_v})/\mathrm{S_d})^2 \quad ---(2)$$

$$R_d = \frac{||d_w \bullet d_v||}{|d_w||\bullet||d_v||} - 1 \quad ---(3)$$

모든 보상은 음의 값으로 주어지며 최댓값은 0이다. 개체가 제한 속도를 준수하는 것을 장려하기 위해 속도에 대한 보상 함수인 수식 (2)는 규정 속도(S_d)와 현재 개체의 속도(S_v) 사이의 차이를 정규화 한 후 제곱한 값의 음수(R_s)를 속도에 대

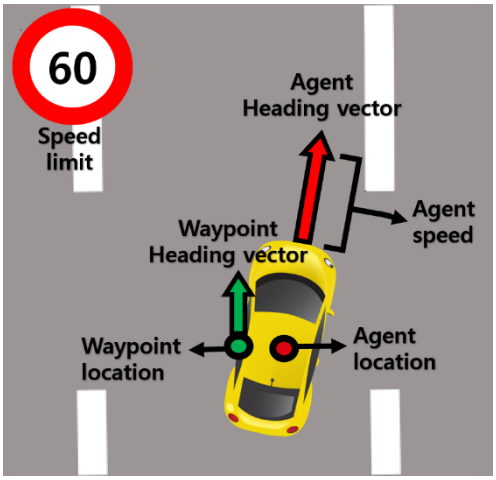

그림 3. 개체의 상태

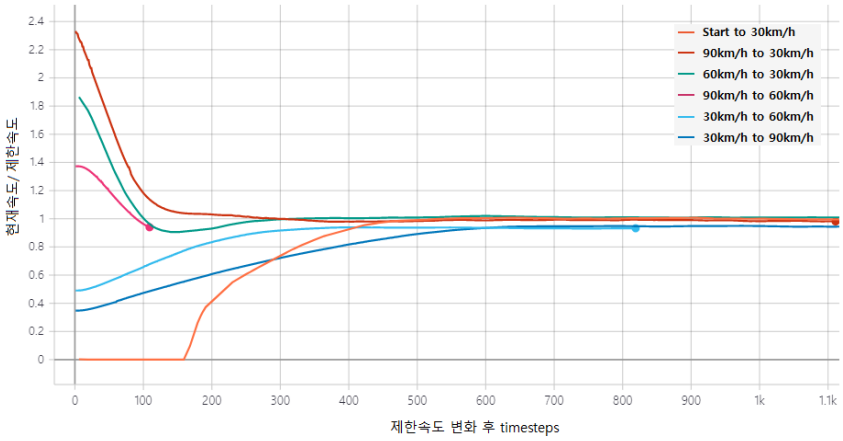

그림 2. 제한속도 변화 후 개체의 속도 변화

한 보상 함수로 사용한다. 회전 구간을 고려한 수식 (3)의 보상함수(R_d)는 회전 구간에서의 안전을 고려하여 제한 속도 보다 더 낮은 속도로 주 행하는 것을 장려하기 위해 차량의 현재 방향 벡터(d_v)와 차선의 방향 벡터(d_w) 사이의 코사인 유사도를 보상으로 사용한다. 이때, 보상 값의 최대가 0이 되게 하기위해 -1 만큼 평행 이동하였다. 최종 보상 함수(R)는 두 보상함수(R_s, R_d)의 합인 수식 (1)로 정의하였다.

PPO는 Conservative Policy Iteration 방법 중 하나로 학습 과정에서 최소한 성능이 떨어지지 않도록 갱신시 정책 모델의 변화 정도를 제한하는 방법 중 하나이다. Advantage는 Policy Gradient에 대한 un biased estimator로서 가장 낮은 분산을 가지고 있으나 실제 값을 알 수 없기 때문에 별도의 estimator를 통해 추정해야 한다.[4]. PPO에서는 이를 위해 Truncated GAE를 사용한다. 하지만 이 논문 에서는 에피소드가 끝나지 않은 상황에서 정책에 대한 갱신을 수행하는 경우 갱신을 하는 동안 Carla 시뮬레이터의 server 측에서는 이전에 개체가 수행한 동작을 반복하여 다음에 얻어지는 표본의 보상이 정확해지지 않는다는 문제가 발생한다. 따라서 정책이 에피소드 별로 갱신되도록 하였다. 이는 GAE에서 λ의 값이 0인 경우에 해당하므로 state value function $V(s)$에 독립적이게 되는 반면 분산이 커진다는 문제점을 가지고 있다.

Ⅲ. 법정 제한 속도 준수를 위한 심층 강화학습 방법

모든 시뮬레이션과 학습은 Carla simulator에서 제공하는 Town01 환경(그림1)에서 진행하였다. Carla의 Town01은 약 300m X 400m의 맵으로 외곽 도로와 중심부의 도로가 존재한다. 도로의 폭은 평균적으로 2m이고 왕복 2차선으로 이루어져 있으며, 실제 도로처럼 폭이나 길이가 일정하지 않다.

모든 도로의 중앙에는 위치 좌표, 진행방향, 도로의 번호등의 정보를 가지는 waypoint가 존재하며, 차량이 waypoint의 위치와 진행 방향을 따라 가는 것을 안전한 주행으로 간주하였다.

Town01의 규정속도는 총 3가지의 규정속도(30km/h, 60km/h, 90km/h)가 존재한다. 외곽 직선 주행 구간에서의 규정속도는 60km/h 또는 90km/h이며, 핸들의 조작이 필요한 교차로, 커브 구간과 중심부의 도로의 규정속도는 30km/h이다.

시뮬레이션 결과 심층 강화 학습으로 학습한 개체가 규정 속도가 변화하는 경우에도 안정적으로 규정 속도를 준수하는 것을 확인하였다. 그림 3은 제한 속도 변화 후 개체의 속도 변화를 나타낸다. x 축은 제한 속도 변화 후의 time step 수를 나타내며 y축은 개체의 현재 속도를 규정 속도로 나눈 값이다. 따라서 x=0 인 시점에서 제한 속도의 변화가 일어나고, y=1 일 때가 차량의 속도와 제한속도가 동일한 가장 바람 직한 상태이다. 그래프의 각 선은 규정 속도 변화에 따라 구분되는 개체의 속도 제어 양상을 보여준다. 처음에 y 값이 1보다 큰 것은 이전 도로의 규정 속도가 현재 규정 속도 보다 큰 값임을 의미하며 1보다 작은 것은 이전 도로의 규정 속도가 현재 규정 속도 보다 작은 값임을 의미한다.

그래프를 보면 약 150 timestep 정도 y 값이 0인 선이 존재하는데 이는 시뮬레이션 에피소드를 시작했을 때 개체의 속도가 0인 상태를 나타낸다. 전체적으로 봤을 때 규정 속도가 60km/h에서 30km /h로 변한 경우나 90km/h에서 30km/h로 변한 경우에는 y 값이 1에 거의 수렴하는 것을 알 수 있다. 반면 규정 속도가 60km/h로 변한 경우나 90km/h로 변한 경우에는 y 값이 약 0.9로 1보다 작음을 확인할 수 있다. 이는 개체가 규정 속도 보다 낮은 속도로 주행함을 의미한다. 특히 90km/h에서 60km/h로 규정 속도가 변하는 경우 에는 약 100 timestep 만에 60km/h에서 다른 규정 속도로 변하기 때문에 규정 속도에 맞추기가 어려움에도 불구하고 규정 속도에 맞게 주행하는 것을 확인하였다. 본 연구의 학습 차량의 운행 동영상은 [5]에서 확인할 수 있다.

Ⅳ. 결론

본 논문은 PPO 알고리즘을 기반으로 학습시킨 자율 주행 차량의 법정 제한 속도 준수에 관한 연구를 수행하였다. 충돌 센서 외에 다른 센서나 이미지를 사용하지 않고, 차량의 관측 가능한 상태인 위치, 진행방향, 속도를 이용하여, 법정 제한 속도를 준수하며 주행하는 것을 확인할 수 있었다. 다만 일부 구간에서 규정 속도를 약간 초과하여 주행하는 것을 확인하였고 이는 현재의 보상함수가 제한속도와의 차이만 중요시하고, 초과하는 경우에 대한 처벌이 포함되어 있지 않기 때문인 것을 확인하였다. 추후 연구에서 처벌함수를 추가하여 본 연구를 개선할 예정이다. 이러한 한계점에도 불구하고 본 연구는 Lidar센서와 같은 고가의 장비 없이 GPS와 네비게이션으로부터 얻을 수 있는 기본적인 정보만 사용해서 자율주행차량이 법정 제한 속도를 준수할 수 있도록 설계할 수 있었다.

Acknowledgement

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

참 고 문 헌

[1] 임재경, 한상진, 엄기종, 이해선, 이선영, 이혜진, “도시부 제한속도 감속(5030)에 따른 교통영향 연구,” 한국 교통 연구원, 국토 교통부, 2017년 11월.

[2] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov, “Proximal Policy Optimization Algorithms,” arXiv preprint arXiv:1707.06347, 2017.

[3] Mümin Tolga Emirler, Esmail Meriç Can Uygan, Bilin Aksun Güvenç, Levent Güvenç, "Robust PID steering control in Parameter Space for highly automated driving," International Journal of Vehicular Technology, 2014.

[4] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan and Pieter Abbeel, “High-Dimensional Continuous Control Using Generalized Advantage Estimation,” International Conference of Learning Representations (ICLR), 2016.

[5] https://drive.google.com/file/d/17qOYIVt1A4KlduVWj6X-A4R497DamLZc/view?usp=sharing