LLM을 이용한 강화학습기반 교차로 신호 제어

최수정[1], 임유진[2]
[1]숙명여자대학교 IT공학과 석사과정
[2]숙명여자대학교 인공지능공학부 교수
suzzang77@sookmyung.ac.kr, yujin91@sookmyung.ac.kr

Reinforcement Learning-Based Traffic Signal Control Using Large Language Models

SuJeong Choi[1], Yujin Lim[2]
[1]Dept. of IT Engineering, Sookmyung Women's University
[2]Division of Artificial Intelligence Engineering, Sookmyung Women's University

요 약

교차로 신호 제어는 스마트 교통 시스템에서 중요한 역할을 하며, 강화학습(RL)을 기반으로 한 연구가 많이 진행되고 있다. 그러나 RL에서 보상 함수 내 요소 간 가중치를 적절하게 설계하는 것은 매우 어려운 과제이다. 최근 LLM을 활용해 데이터를 분석하고 의사 결정을 보완하는 연구가 주목받고 있으며, 이를 RL에 적용해 이러한 문제를 해결하려는 시도도 진행 중이다. 본 논문에서는 실시간으로 변화하는 교차로 내 효율적인 교통 제어를 위해 LLM을 이용하여 RL 보상 함수의 가중치를 실시간으로 조절함으로써 교통신호를 효율적으로 제어하는 알고리즘을 제안한다. 그리고 초기 가중치 설정과 가중치 조절 주기에 대한 실험 결과를 비교함으로써 제안 기법의 성능을 분석한다.

1. 서론

교차로 신호 제어는 도시 교통 관리에서 중요한 역할을 하며, 차량 흐름과 이동 시간을 크게 좌우한다. 기존 방식은 고정된 규칙을 이용해 실시간 상황에 유연하게 대응하지 못하는 한계가 있다. 이를 해결하기 위해 강화학습(RL)이 주목받고 있으며, 실시간 데이터 기반 교통신호 제어를 통해 교통 흐름을 개선할 수 있다. 그러나 RL의 성능은 보상함수 설계에 크게 의존하며, 복잡한 교통 패턴을 반영하는 적절한 보상함수 설계는 여전히 어려운 과제이다[1].

최근에는 대형 언어 모델(Large Language Models, LLM)이 자연어 처리뿐만 아니라 다양한 의사 결정 및 데이터 분석 영역에서 뛰어난 성능을 보여주고 있다. 이와 함께 RL을 개선하기 위한 LLM 활용 연구도 활발히 진행되고 있다. 본 논문은 보상 함수 설계에 LLM을 활용해 실시간 교통 상황에 유연하게 대응할 수 있는 강화학습 기반 신호 제어 알고리즘을 제안한다.

2. 관련 연구

RL 기반 교차로 신호 제어에 대한 연구로는 [1]이 있으며, 이 연구에서는 효과적인 마르코프 결정과정(MDP)을 설계하였다. 그러나 보상함수를 구성할 때 각 요소의 가중치에 따라 결과가 달라져 적절한 가중치 설정에 어려움이 있다고 언급되었다. 따라서 최근에는 이러한 RL의 한계점을 보완하기 위해 LLM이 사용되고 있는데, 교차로 신호 제어 연구에서도 실시간으로 변화하는 교통 상황을 기반으로 좀 더 정확한 신호 결정을 내리기 위해 LLM을 사용한 기법들이 많이 제안되고 있다. [2]는 LLM 기반 교차로 신호 제어에 관한 연구로, Light GPT를 사용하여 교차로 교통 상황을 분석하고 그에 맞는 신호를 결정하는 방식을 제안하였다. 이 연구는 신호 제어 에이전트로 RL을 사용하는 대신 LLM을 활용했으나, LLM이 RL처럼 목표를 달성하기 위한 최적의 정책을 학습하는 기능을 완전히 대체하기는 어려웠다. 따라서 LLM을 에이전트로 사용하는 것은 교차로 신호 제어 최적화에 한계가 있다. [3]은 RL 기반 교차로 신호 제어에 LLM을 활용하여 교차로 신호 제어 에이전트로서 RL이 결정한 신호에 대한 피드백을 제공하고, 그 피드백을 바탕으로 신호를 다시 결정하는 방법을 제시하였다. 즉, 해당 연구에서 LLM은 RL이 내린 신호를 평가하고 적절한 신호를 도출하는 역할을 한다. 그러나 이 연구는 사전에 학습된 RL 모델에 단순히 LLM을 통해 행동

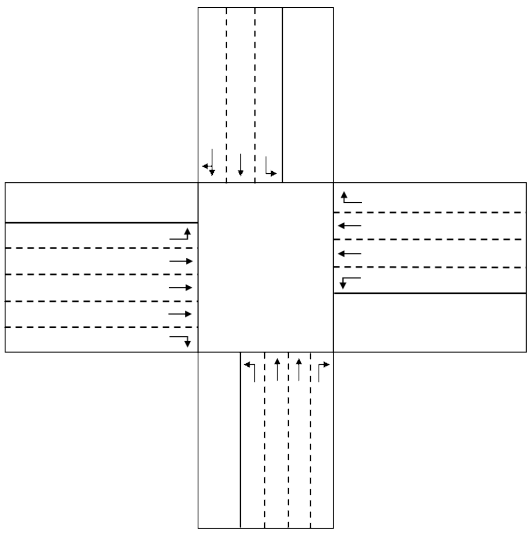
(action)에 대한 피드백만 제공하는 방식으로, RL의 정책을 직접적으로 개선하기는 어려운 한계가 있다.

따라서 본 논문은 기존 연구의 한계점을 해결하기 위해 LLM을 RL 기반 교차로 신호 제어 시스템에 통합하여, 보상함수 설계의 어려움을 극복하는 새로운 알고리즘을 제안한다. LLM을 활용해 교차로 교통 상황을 분석하고 이를 통해 실시간으로 보상함수의 가중치를 도출하고자 한다.

3. 시스템 모델

3-1. 환경 설정

실험 환경은 단일 교차로이며 (그림 1)과 같고 각 차선별로 직진, 좌회전, 우회전이 가능하며 우회전은 신호 관계 없이 항상 가능하다고 가정했다.

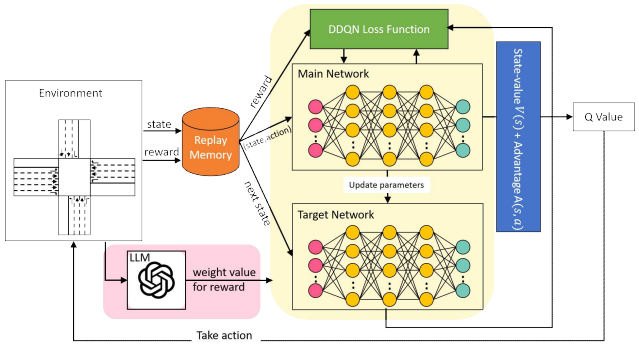


(그림 1) 단일 교차로 환경

실험을 위한 교차로 트래픽 데이터 세트는 SUMO(Simulation of Urban Mobility)를 통해 발생시켰다[4]. 이동하는 차량의 크기는 5.00m이고 대기중인 차량 간의 간격은 2.50m로 설정하였다. 또한 각 차선의 넓이는 3.20m로 모두 일정하며 도로의 길이는 약 3km이다.

3-2. 제안 알고리즘

에이전트의 최종적인 목표는 교차로의 평균 여행 시간(average travel time)을 최소화하는 것이다. 여행 시간은 교차로에 연결된 차선들에 진입한 순간부터 교차로를 빠져나갈 때까지의 시간이다. 본 논문에서는 Dueling Double DQN(D3QN) 알고리즘을 사용하여 교차로 신호를 제어하며, 보상 함수의 가중치 조절에는 LLM을 활용한다. 제안하는 모델은 D3QN-LLM으로 명시하고, 프레임워크는 (그림 2)와 같다.



(그림 2) D3QN-LLM 프레임워크

D3QN은 DQN을 개선한 알고리즘으로, Double DQN[5]과 Dueling DQN[6]을 결합한 버전이다. Double DQN은 DQN이 과대평가 문제로 인해 최적의 action을 학습하지 못하는 문제를 해결하기 위해 개발되었으며, Dueling DQN은 Q-value의 노이즈에 강하여 보다 더 안정적인 학습이 가능하다. 본 논문에서는 더 효과적인 학습을 위해 이 두 가지 기법을 결합한 D3QN을 선택하였다.

본 논문에서 정의한 상태(S_t), 행동(A_t), 보상(R_t)의 정의는 다음과 같다. 상태 S_t는 시간 t에서 교차로의 각 방향(동,서,남,북)에 대해 직진(straight), 좌회전(left) 차선을 구분하여 존재하는 차량 수와 현재 할당된 신호(P_t)로 나타낸다. 존재하는 차량 수에 대해 동쪽 직진 차선은 $N_{east\_s,t}$, 좌회전 차선은 $N_{east_l,t}$로 나타내고 서쪽 직진 차선은 $N_{west\_s,t}$, 좌회전 차선은 $N_{west_l,t}$이다. 그리고 남쪽 직진 차선은 $N_{south\_s,t}$, 좌회전 차선은 $N_{south_l,t}$이고, 북쪽 직진 차선은 $N_{north\_s,t}$, 좌회전 차선은 $N_{north_l,t}$로 표현한다. 따라서 상태 S_t은 식(1)과 같이 표현할 수 있다.

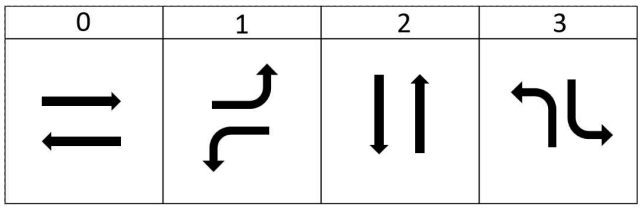
$$S_t = \{N_{north\_s,t}, N_{north\_l,t}, N_{south\_s,t}, N_{south\_l,t}, N_{west\_s,t}, N_{west\_l,t}, N_{east\_s,t}, N_{east\_l,t}, P_t\} \quad (1)$$

에이전트는 현재 상태 S_t를 고려하여 행동 A_t를 선택하고 그 행동을 수행한다. 행동은 할당 가능한 녹색 신호를 나타내고 (그림 3)처럼 행동 $A_t = \{0,1,2,3\}$로 정의했다.

에이전트는 행동 A_t를 수행함으로써 보상 R_t을 얻는다. 본 논문의 목표인 평균 여행 시간을 최소화하기 위해 구성한 보상 함수는 식(2)와 같다. [1]을

참고하여 travel time에 영향을 주는 요소인 대기시간(waiting time)과 대기 차량 수(queue length)로 구성했다. 대기시간은 정지 신호에 차량이 정지한 순간부터 출발할 때까지 시간이다. 그리고 대기 차량 수는 정지 신호에 멈춰 있는 차량 수를 의미한다. 교차로에 연결된 도로는 3km이므로 3km 범위까지의 차선에 존재하는 모든 차량의 대기시간, 그리고 대기 차량 수를 측정한다. 그리고 해당 지점부터 차량의 여행 시간을 측정한다.

교차로의 공평한 신호 분배를 위해 표준편차(std)를 사용하여 각각 T_{std}, Ql_{std}로 나타냈다. T_{std}는 교차로 방향별 차량의 대기시간의 표준편차, Ql_{std}는 교차로 방향별 대기 차량 수의 표준편차를 의미한다. α는 가중치로서 0과 1 사잇값이고 교통 상황에 따라 달라질 수 있다.



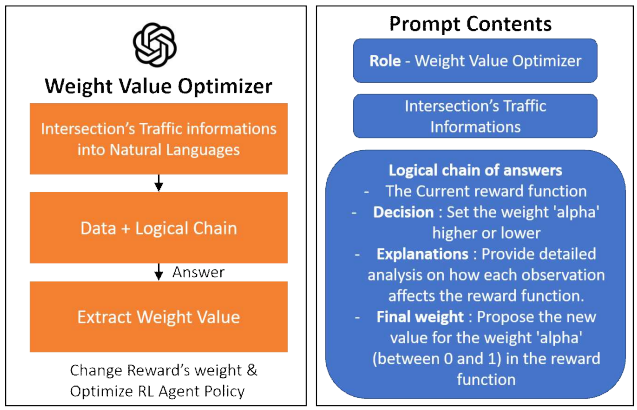
(그림 3) 행동공간 (action space)

$$R_t = -\{\alpha \cdot T_{std} + (1-\alpha) \cdot Ql_{std}\} \quad (2)$$

보다 효과적인 학습을 위해서는 적절한 가중치 α를 찾아야 하지만, 이를 사람이 경험적으로 설정하는 것은 매우 어렵고 에이전트가 최적의 정책을 학습하는 데 한계가 있다. 이를 해결하기 위해 LLM을 활용해 action 후 교통 상황을 분석하고, 그에 따라 최적화된 가중치를 결정하고자 한다.

성능과 모델 크기를 고려해 OpenAI의 GPT-4o mini를 LLM 모델로 선택했다. 이 모델은 경량화된 GPT-4o로, 성능을 유지하면서도 적은 자원을 사용하도록 설계되었다. 또한, LangChain 프레임워크를 통해 LLM 모델을 쉽게 불러와 프롬프트 엔지니어링 등을 간편하게 활용할 수 있다. Weight Value Optimizer의 구조 및 프롬프트는 [3]을 참고해 설계되었으며, (그림 4)에서 확인할 수 있다.

(그림 2)에서 볼 수 있듯이, RL 에이전트는 action을 취한 후 교차로 교통 상황을 LLM에 전달하여 실시간으로 가중치를 도출하고, 이를 바탕으로 reward를 계산한다. 이를 통해 RL 에이전트는 더 정교한 보상 구조를 학습하며, 최적의 신호 제어 정책을 도출해 교통 흐름을 개선할 수 있다.



(그림 4) LLM 구조와 프롬프트 내용

4. 실험 결과 및 분석

가중치 조절 주기에 따른 성능을 비교하기 위해 초기 학습 시작 가중치를 0.2, 0.5, 0.8로 각각 설정하고 가중치 조절 주기는 에피소드 단위로 1, 10, 50, 100으로 설정했다. 성능 비교 지표는 교차로의 평균 여행 시간(Average Travel Time, ATT), 평균 대기 차량 수(Average Queue Length, AQL), 평균 대기 시간(Average Waiting Time, AWT)이다.

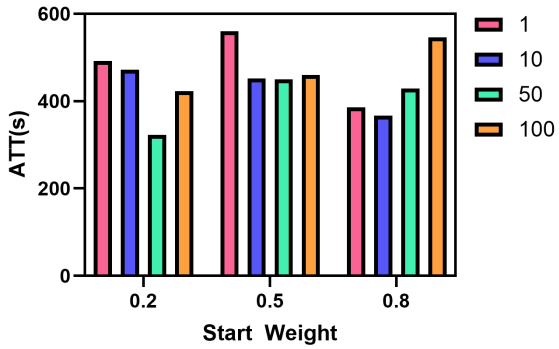
(그림 5)는 가중치 조절 주기에 따른 초기 학습 가중치 별 ATT를 나타낸다. 초기 가중치가 0.2이고 조절 주기가 50일 때 ATT는 323초로 가장 짧았다. 그리고 초기 가중치가 0.5이고 조절 주기가 1일 때 560초로 가장 길었고 초기 가중치가 0.8이고 주기가 100일 때 546초로 그 다음으로 길었다. 전체적으로 보면 초기 가중치가 0.2, 0.8일때는 ATT 값의 범위가 매우 넓지만 0.5일 때 주기가 1일 때를 제외하곤 약 440초대로 중간 정도의 값을 유지한다.

(그림 6)은 가중치 조절 주기에 따른 초기 학습 가중치 별 AQL을 나타낸다. 초기 가중치가 0.5이고 주기가 10, 50일 때 AQL이 4로 가장 짧았다. 그리고 초기 가중치가 0.2, 주기가 1일 때 AQL이 6이고 초기 가중치가 0.8일 때 주기 1을 제외하곤 AQL이 6으로 가장 길었다. 전반적으로 초기 가중치가 0.5일 때 낮은 AQL값을 보임을 알 수 있다.

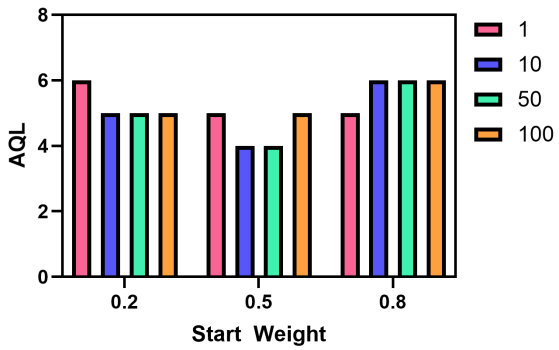
(그림 7)은 가중치 조절 주기에 따른 초기 학습 가중치 별 AWT를 나타낸다. 초기 가중치가 0.2이고 주기가 1일 때 351초로 가장 길었다. 그리고 초기 가중치가 0.5이고 주기가 50일 때 197초로 가장 짧았다.

ATT, AQL, AWT는 상호 연관된 지표들이므로,

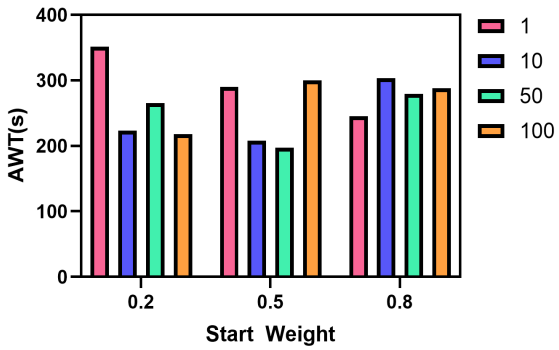
각각의 성능뿐만 아니라 이들을 함께 분석해야 전체적인 성능을 정확하게 평가할 수 있다. 따라서 초기 학습에서 가중치를 0.5로 설정한 경우, AQL의 증감에 따라 ATT와 AWT도 동일한 패턴으로 증감하는 양상을 보여주며, 가장 안정적인 결과를 나타냈다. 특히, 가중치 조절 주기가 10 또는 50일 때가 가장 적절한 것으로 판단된다.



(그림 5) Average Travel Time(ATT)



(그림 6) Average Queue Length(AQL)



(그림 7) Average Waiting Time(AWT)

5. 결론

본 연구에서는 강화학습기반 교차로 신호 제어에 LLM을 이용하여 실시간으로 보상함수의 가중치를 조절하는 알고리즘을 제안했다. 이를 통해 RL의 효율적인 정책 학습을 촉진하고 목표를 위해 reward를 최대화하고자 했다. 실험 결과를 통해 초기 학습 가중치를 0.5로 설정하고 가중치 조절 주기가 10 또는 50일 때 가장 안정적인 결과가 나오는 것을 알 수 있었다.

이를 통해, 추후 연구에서는 교통량 부하를 조정하여 실제와 유사한 환경에서 실험을 진행할 계획이다. 또한, 다른 교차로 신호 제어 알고리즘과 비교하여 LLM을 활용한 실시간 가중치 조절이 얼마나 효과적인지 평가할 것이다.

사사문구

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음 (IITP-2024-RS-2022-00156299)

참고문헌

[1] G. Zheng, X. Zang, N. Xu, H. Wei, Z. Yu, V. Gayah, et al., "Diagnosing reinforcement learning for traffic signal control," arXiv:1905.04716, 2019.

[2] S. Lai, Z. Xu, W. Zhang, H. Liu and H. Xiong, "Large language models as traffic signal control agents: Capacity and opportunity," arXiv:2312.16044, 2023.

[3] A. Pang, M. Wang, M. O. Pun, C. S. Chen, and X. Xiong, "iLLM-TSC: Integration reinforcement learning and large language model for traffic signal control policy improvement," arXiv:2407.06025, 2024.

[4] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, P. Flotterod, R. Hilbrich, L. Lucken, J. Rummel, P. Wagner and E. Wiessner, "Microscopic traffic simulation using sumo," International Conference on Intelligent Transportation Systems (ITSC), 2018.

[5] H. Van Hasselt, A. Guez and D. Silver, "Deep reinforcement learning with double q-learning," AAAI Conference on Artificial Intelligence, 2016.

[6] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot and N. Freitas, "Dueling network architectures for deep reinforcement learning," International Confrence on Machine Learning (ICML), 2016.