Journal of Next-generation Convergence Information Services Technology Vol.15, No.1, February (2026), pp.73-81
ISSN 2384-101X(p) 2672-1163(e) *http://dx.doi.org/10.29056/jncist.2026.02.07*

레이스 트랙 환경에서 심층 강화 학습 에이전트의 성능 향상을 위한 자동 커리큘럼 학습 적용 연구

A Study on Applying Automated Curriculum Learning to Improve the Performance of Deep Reinforcement Learning Agents in a Race Track Environment

이동철[1]

Dongcheul Lee[1]

요 약

복잡한 자율주행 환경에서 에이전트를 처음부터 학습시키는 것은 불안정한 학습 수렴 문제에 직면하기 쉽다. 본 연구는 이 문제를 해결하기 위해 자동화된 커리큘럼 학습을 적용한 방법을 제안한다. 초기 학습에서 과도한 충돌과 부정적 보상으로 인한 지역해 수렴 문제를 해결하기 위해, 학습 시 Teacher-Student Framework를 사용하여 환경 난이도의 핵심 변수인 상대 차량 최대 수를 동적으로 조절하였다. Teacher는 최근 에피소드의 성공률을 기준으로 차량 수를 조절하고, cooldown 및 interleaving 기법을 이용하여 학습의 안정성을 높였다. 실험은 highway-env의 racetrack-v0 환경에서 실행하였다. 그 결과 제안된 ACL 방식이 기준 모델 대비 보상, 에피소드 길이, 생존 시간, 충돌 횟수에서 큰 폭의 향상을 보였다. 또한 학습 시에도 기준 모델 대비 안정적으로 높은 보상 및 생존 시간에 수렴하는 것을 볼 수 있었다. 이는 제안하는 방법이 초기에 기본적인 자동차 주행 방법을 학습한 뒤, 복잡한 환경에서 회피하는 전략까지 순차적으로 학습하도록 유도하여 지역해 문제를 완화하고 더 우수한 정책을 획득하게 했음을 보여준다.

핵심어 : 강화 학습, 커리큘럼 학습, 자율 주행, 교사-학생 프레임워크

Abstract

Training an agent from scratch in a complex autonomous driving environment often results in unstable convergence during learning. To address this challenge, this study introduces an approach based on automated curriculum learning. In order to avoid early local optima caused by frequent collisions and negative rewards, the method uses a Teacher-Student Framework that adjusts the maximum number of surrounding vehicles, which represents the core factor in determining task difficulty. The Teacher adjusts this number according to the recent success rate of episodes and improves learning stability through cooldown and interleaving techniques. Experiments were conducted in the racetrack-v0 environment of

1 Department Multimedia Engineering, Hannam University, Daejeon, Korea [Professor]
e-mail: jackdclee@hnu.kr

* 이 논문은 2025학년도 한남대학교 학술연구비 지원에 의하여 연구되었음.

Received(November 21, 2025), Review Result(1st: December 15, 2025), Accepted(February 13, 2026), Published(February 28, 2026)

highway-env. The results show that the proposed ACL method provides substantial improvements over the baseline in reward, episode length, survival time, and collision count. The learning process also converged to consistently higher rewards and longer survival times compared to the baseline. These results indicate that the proposed approach helps the agent gradually learn basic driving skills first and then acquire more advanced avoidance strategies in complex traffic, reducing the likelihood of falling into local optima and leading to better overall policies.

Keyword : Reinforcement learning, Curriculum learning, Autonomous driving, Teacher-Student Framework

1. 서론

미래 교통 시스템의 패러다임을 새롭게 열 핵심 기술로 주목받는 자율주행 기술은 사용자의 안전성과 편의성을 증대시키고 교통 효율을 극대화할 수 있다는 장점이 있다. 완전한 자율주행을 실현하기 위한 가장 큰 난관 중 하나는 수많은 변수가 실시간으로 상호 작용하는 복잡한 환경에서 어떻게 동작할지 행동을 결정하는 일이다. 이러한 복잡한 환경에서 에이전트가 특정 행동을 결정하는 것은 기존의 규칙 기반 시스템으로는 불가능하며 환경과 상호작용을 통해 시행착오를 거쳐 최적 행동을 결정하는 강화학습이 그 해결책으로 제시되었다. 특히 심층 신경망과 결합한 심층 강화학습 (DRL, Deep Reinforcement Learning)은 복잡한 변수들을 입력받아 주행 행동을 결정하는데 뛰어난 성능을 보였다 [1].

DRL의 잠재력에도 불구하고 실제와 비슷한 복잡한 시뮬레이션 환경에서 학습 에이전트를 성공적으로 학습시키기는 여전히 어려운 과제이다. 전통적인 DRL은 시뮬레이션 환경의 최종 목표를 처음부터 고정하여 에이전트를 학습시킨다. 그러나 이렇게 고정된 최종 목표를 학습 초기부터 사용한다면 학습 시 탐색을 소극적으로 하도록 유도하여 결국은 학습을 진행하는 데 어려움을 초래한다. 왜냐하면, 에이전트가 트랙을 유지하는 방법을 학습하기 전에 빈번한 충돌을 경험하여 부정적인 보상을 너무 많이 받기 때문이다. 이러한 과도한 부정적 보상 상황에서는 에이전트는 매우 소극적으로 반응하여 매우 느린 속도로만 운행하도록 학습하게 된다. 이렇게 하면 큰 부정적 보상은 피할 수 있지만 장기적으로는 적당한 속도로 충돌은 피하면서 오래 달리는 본질적인 목표를 달성하기 어렵게 된다. 특히 학습 초기 단계에서 복잡한 문제에 너무 많이 노출되어 과도한 실패를 경험하게 되면 유의미한 학습을 하지 못하고 지역 최적해(Local Optima)에 수렴하거나 학습을 실패하게 된다 [2].

이러한 한계에 대한 효과적인 해결책으로 커리큘럼 학습(Curriculum Learning, CL)을 고려할 수 있다 [3]. 커리큘럼 학습은 에이전트에게 처음부터 최종 목표를 제시하는 것이 아니라 난이도가 점진적으로 증가하도록 일련의 중간 목표들을 순차적으로 제시하여 학습시키는 것을 말한다. 이를 통해 쉬운 단계에서는 기초 기술을 학습하고 이를 바탕으로 어려운 단계의 더 복잡한 기술을 습득할 수 있도록 돕는다.

본 연구에서는 Farama Foundation에서 제공하는 highway-env [4]의 Racetrack-v0 환경에서 효과적

인 커리큘럼 학습 전략을 통해 에이전트가 얼마나 잘 학습하는지 알아볼 것이다. 이 환경은 다수의 자동차가 존재하는 고속 도로 주행 시나리오를 모델링한다. 이러한 환경에서 DRL 에이전트의 목표는 단순히 트랙을 따라 주행하는 것을 넘어 주변 차량의 움직임을 예측하고 충돌을 회피하며 안정적으로 오랫동안 주행하는 것이다. 본 논문에서 제시하는 커리큘럼 학습을 한 에이전트의 성능을 평가하기 위해 커리큘럼 학습을 하지 않은 에이전트와 학습 결과를 비교 평가할 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 자율주행을 위한 DRL 및 커리큘럼 학습에 대한 기존 연구들을 고찰한다. 3장에서는 본 연구에서 사용한 시뮬레이션 환경 및 DRL 에이전트, 그리고 제안하는 커리큘럼 설계에 대해 기술한다. 4장에서는 실험 설정, 평가 지표, 그리고 제안하는 방법과 기준 모델의 성능 비교 평가를 분석한다. 마지막으로 5장에서는 연구 결과를 요약하고 시사점과 향후 연구 방향을 논한다.

2. 관련 연구

커리큘럼 학습은 에이전트가 최종 과제 목표부터 학습하는 것이 아니라 현재 능력에 맞춰 과제의 난이도를 점점 높여나가면서 학습하는 것이다. 이를 통해 희소 보상 문제(Sparse reward problem) [5]를 줄이고, 비효율적인 탐색을 완화하여 더 나은 학습 결과와 빠른 수렴 속도를 달성하는데 도움을 준다. CL은 크게 수동 커리큘럼 학습(Manual CL, MCL)과 자동 커리큘럼 학습(Automatic CL, ACL)으로 나뉠 수 있다.

2.1 MCL

대부분의 초기 커리큘럼 학습 연구들은 MCL 방식을 택했다. MCL은 학습 단계를 여러 개로 나누어서 각 단계의 복잡성을 점진적으로 증가시키는 학습 방법이다. 단계의 전환은 주로 전문가가 정한 휴리스틱에 의해 결정된다. [6]은 자율 주행 학습을 위해 각 학습 단계마다 교통 시나리오의 복잡성을 점진적으로 증가시키는 방법을 제안했다. 예를 들어 초기 단계에는 다른 차량이나 보행자가 없는 환경에서 학습하고, 다음 단계에서 정적인 장애물을, 그 다음 단계에서는 동적인 차량을 추가하는 방식이다. 이 방식은 문제 구조가 명확하고 전문가의 개입이 용이한 환경에서는 우수한 성능을 보였으나, 전문가가 비즈니스 도메인 지식을 활용하여 커리큘럼을 설계해야 하므로 일반화가 어렵고 새로운 환경에 대한 확장성이 부족하다는 한계를 가진다. 또한 직관에 의존한 난이도 조정이 부정확할 경우 오히려 학습 성능이 더 저하될 수도 있다.

2.2 ACL

이러한 MCL의 한계를 해결하기 위해 ACL은 알고리즘이 현재 성능을 기준으로 동적으로 학습

순서를 정한다. [7]는 에이전트의 학습 진척도를 기준으로 과제의 목표를 자동으로 정하는 Teacher-Student Framework를 사용했다. 여기서 Teacher는 학습할 과제를 선택하는 역할을 하는 메타 모델이며, Student는 실제 환경에서 과제를 학습하는 주 모델이다. 이 방법에서 Teacher는 Student의 학습 향상률을 가장 증대시키는 적절한 난이도의 과제를 동적으로 선택하여 Student가 학습 정체에 빠지는 것을 방지하고 내재적 동기를 부여하여 지속적인 학습을 유도하였다. [8]은 Generative Adversarial Network(GAN)을 이용하여 에이전트가 도전할 만한 새로운 목표 상태를 만들어내는 방식으로 커리큘럼을 구축하였다. 이를 바탕으로 별도의 보상 함수 설계 없이 탐험되지 않은 상태 공간으로 에이전트를 점진적으로 유도하여 학습 효율을 향상시켰다. [9]은 Student의 gradient norm을 Teacher의 보상 신호로 사용하는 보상 기반 적응형 방식을 사용하여 모델이 가장 학습 효과가 큰 데이터 분포로 스스로 이동하도록 하였다. 이 방식은 보상의 변화가 적더라도 gradient가 크다면 미세한 학습 진전을 더 빨리 포착하여 적절한 난이도로 안내할 수 있다. 이렇게 ACL을 활용하면 전문가의 개입을 줄일 수 있으므로 환경 변화에 빠르게 대처할 수 있고 자동화된 대규모 학습이 가능하게 된다. 그러나 환경이 복잡해질수록 알고리즘 설계 복잡도와 계산 비용이 증가하는 단점이 있다.

3. 연구 방법

본 연구는 Farama Foundation에서 개발한 자율 주행 시뮬레이션 환경 모음인 highway-env 내의 racetrack-v0을 사용한다. 이 환경은 에이전트가 폐쇄된 트랙을 따라 주행하면서 다른 차량과 충돌을 피해야 하는 과제를 제공한다. 따라서 에이전트는 트랙을 이탈하지 않고 다른 차들과 충돌하지 않으면서 최대한 오랫동안 멀리 주행하는 것이 주된 목표이다. 에이전트의 기본 알고리즘으로는 PPO(Proximal Policy Optimization)를 사용하였다. PPO는 정책 기반의 알고리즘으로 이전 정책과 새로운 정책의 차이가 너무 커지지 않도록 클리핑을 통해 업데이트의 크기를 제한하는 것이 특징이다. 이러한 특징으로 다른 알고리즘에 비해 하이퍼파라미터 설정에 덜 민감하고 학습 과정이 안정적이기 때문에 복잡한 연속 제어 문제를 해결하는데 널리 사용된다 [10].

에이전트가 사용한 보상 함수 중 차선 중심 유지를 위한 보상함수는 다음과 같다. l은 차선 중심과 차량 중심의 거리를 뜻하며 k는 그 가중치이다.

$$R_{ce} = \frac{1}{1 + kl^2} \tag{1}$$

행동 벡터에 대한 보상함수는 다음과 같다. 에이전트의 조향 벡터 a를 크게할수록 패널티가 증가한다.

$$R_{ac} = \| a \|_2 \tag{2}$$

차선 중심 유지, 행동 벡터 변화, 충돌 여부에 대한 보상 함수의 가중 합은 다음과 같다. 여기서 w_{ce}, w_{ac}, w_{co}는 각각 차선 유지 가중치, 행동 벡터 가중치, 충돌 가중치를 의미한다. R_{co}는 충돌 여부 {0, 1}을 나타내는 보상이다.

$$S = w_{ce}R_{ce} - w_{ac}R_{ac} - w_{co}R_{co} \tag{3}$$

최종 보상 함수 R은 다음과 같이 정의한다. 여기서 R_{on}은 자동차가 도로 위에 있는지를 뜻하며 도로 위일 경우 1, 도로 밖일 경우 0 값을 지녀 모든 보상을 차단하도록 한다.

$$R = \left(\frac{S + w_{co}}{1 + w_{co}} \right) \bullet R_{on} \tag{4}$$

racetrack-v0에서 다른 차량의 최대 수(other_vehicles_count)는 10대(max_vehicles)로 설정하였다. 이러한 상황에서는 도로에서 다른 차량을 자주 만나게 되므로 에이전트가 다른 차량을 피하도록 학습하는 것이 필수적이다. 따라서 본 연구는 에이전트가 상대 차량의 수에 기반하여 복잡한 다중 차량 환경에 점진적으로 적응하도록 돕는 Teacher-Student Framework 기반 ACL을 설계할 것이다. 기본 아이디어는 에이전트가 다른 차량이 없는 환경에서 트랙을 안정적으로 주행하는 능력을 완벽하게 학습한 후 주변 차량을 피하는 능력을 점진적으로 학습하도록 유도하는 것이다. 쉬운 단계에서는 충돌 회피라는 복잡한 변수의 영향을 최소화하고 에이전트가 트랙을 안정적으로 주행할 수 있도록 하는 것이 목표이다. 상대 차량이 없거나 매우 적은 환경에서 트랙 경계를 인식하고, 곡선 도로에서 벗어나지 않도록 안정적으로 주행하며, 차선의 중앙을 유지하는 능력을 학습한다. 주변 차량이 늘어나는 어려운 단계에서는 단기적으로는 앞선 한 차량을 추월하는 능력을 학습하고, 장기적으로는 앞선 한 차량뿐만 아니라 주변 여러 차량을 고려하여 충돌하지 않도록 주행하는 능력을 학습할 것이다.

Teacher의 역할은 Student가 학습의 효과를 최대한으로 끌어낼 수 있는 최적의 상대 차량 최대 수를 설정한 환경을 제공하는 것이다. Teacher는 Student가 현재 설정된 차량 수에서 학습하는 동안 최근 window_size 개의 에피소드의 성공률을 지속해서 모니터링한다. 성공률이 upper_thresh 값 이상일 경우 차량 수를 max_increment만큼 증가시킨다. 만약 성공률이 lower_thresh 값 더 낮을 때 차량 수를 max_decrement만큼 감소시킨다. 성공률은 충돌 없이 max_survival_time 이상 주행했을 경우 성공, 그렇지 않을 경우 실패로 간주한다. 차량 수를 변경한 직후에는 안정화를 위해 cooldown의 에피소드 만큼 레벨을 변경하지 않으며 학습한다. 또한 catastrophic forgetting을 방지하기 위해 난이도 증가 후에도 주기적으로 과거 난이도를 interleave_ratio 비율로 섞어 학습하도록 하여 안정성을

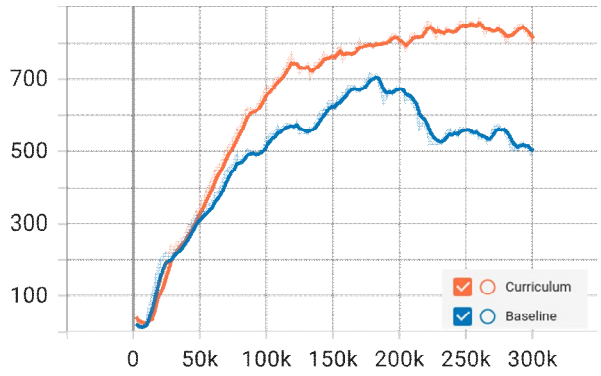
유지하도록 한다 [11].

실제 에이전트 학습은 2단계로 진행된다. 1단계에서는 Teacher 없이 단일 차량으로만 학습하여 안정화될 때까지 학습한다. 2단계에서는 Teacher-Student Framework를 사용하여 차량 수가 max_vehicles+1에 이를 때까지 학습한다.

4. 실험 결과

제안하는 CL 방식의 학습 효과를 검증하기 위해 기준 모델을 설정하였다. 기준 모델은 커리큘럼 학습을 적용하지 않고 학습 시작부터 끝까지 최종 목표 환경과 동일하게 다른 차량의 최대 수를 10대로 설정하였다. 그 외에 PPO 알고리즘, 하이퍼파라미터, 총 학습 시간 등 모든 조건을 동일하게 유지하였다.

[그림 1]은 학습하는 동안 제안하는 CL 방식과 기준 모델의 평균 보상을 비교하는 그래프이다. 제안하는 CL 방식의 경우 비교적 안정적으로 우상향하는 경향을 보인다. 이는 에이전트가 각 단계의 목표를 성공적으로 학습하고 더 어려운 단계에 효과적으로 적응하고 있음을 나타낸다. 또한 학습 후반부로 갈수록 변동폭이 적은 것은 정책이 안정적으로 수렴하고 있음을 보여준다. 그러나 기준 모델은 학습 초반에는 보상이 비교적 빠르게 상승하지만 200k 타임스텝 지점에서 급격히 하락하는 현상을 보인다. 이렇게 학습 후반 부에 변동 폭이 큰 것은 학습하는 동안 정책이 안정되지 못했음을 나타낸다. 이는 에이전트가 우연히 발견했던 불안정한 정책을 유지하지 못하고 충돌만 피하기 위해 소극적인 정책을 택하여 지역해에 빠졌음을 나타낸다.

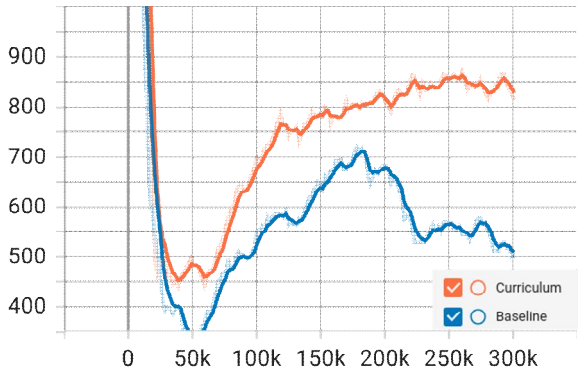


[그림 1] 학습하는 동안 제안하는 CL 방식과 기준 모델의 평균 보상 비교

[Fig. 1] Comparison of average reward during training between the proposed CL method and the baseline

[그림 2]는 학습하는 동안 제안하는 CL 방식과 기준 모델의 평균 에피소드 길이를 비교하는 그래프이다. 에피소드 길이도 보상과 마찬가지로 제안하는 방식은 안정적으로 우상향하는 모습을 보

이고 있으며, 비교 모델은 전반적으로 제안하는 방식보다 낮은 값을 보이며 평균 보상 그래프와 마찬가지로 200k 타임스텝 지점에서 급격히 하락하는 현상을 보인다.



[그림 2] 학습하는 동안 제안하는 CL 방식과 기준 모델의 평균 에피소드 길이 비교

[Fig. 2] Comparison of average episode length during training between the proposed CL method and the baseline

학습이 완료된 두 에이전트의 성능을 평가하기 위해 각각 100회의 테스트 에피소드를 실행하여 평균 보상, 평균 스텝, 평균 생존 시간, 도로를 벗어난 횟수, 충돌 횟수를 측정하였다. 그 결과는 [표 1]과 같다.

[표 1] 기준 모델과 제안하는 CL 방법의 테스트 결과 비교

[Table 1] Comparison of test results between the proposed CL method and the baseline

	Baseline	Curriculum	Change Ratio (%)
Reward	615.0 ± 523.0	869.0 ± 559.0	41%
Steps	621.3 ± 526.6	924.1 ± 572.1	49%
Time	6.0 ± 5.5	9.2 ± 6.6	52%
Off-roads	1.2 ± 2.4	5.6 ± 13.0	379%
Crash	84	61	-27%

제안하는 CL 방식은 에피소드 당 평균 보상, 평균 스텝, 평균 생존 시간에서 모두 비교 모델에 비해 우수한 성능을 보였다. CL 방식의 평균 보상은 869.0으로 기준 모델의 615.0보다 41% 더 높았다($t(197)=3.32$, $p=0.00108$). 또한 평균 스텝도 CL 방식의 경우 924.1로 기준 모델의 621.3보다 49% 높았다($t(196)=3.89$, $p=0.000135$). 평균 생존 시간도 CL 방식은 9.2로 기준 모델의 6.0보다 52% 높았다($t(191)=3.65$, $p=0.000332$). 이는 CL 에이전트가 비교 모델에 비해 훨씬 더 오랫동안 안정적으로 다른 차량과 충돌 없이 주행했음을 의미한다. 이는 총 충돌 횟수에서도 확인할 수 있는데 제안 방법은 61회이지만 기준 모델은 무려 84회 충돌로 27% 차이를 보였다.

에피소드 당 도로를 벗어난 평균 횟수는 CL 방식이 5.6회로 기준 모델의 1.2회 보다 379% 더 많았다. CL 방식이 더 많은 이유는 충돌이 더 적었기 때문에 그만큼 주행시간이 긴 경우가 많아서 도로를 벗어난 경우도 많았던 것으로 판단된다. 기준 모델의 경우 도로를 벗어나기 전에 이미 충돌을하여 에피소드를 종료했기 때문에 낮은 횟수를 유지할 수 있었다. 이는 CL 방식이 주로 완주 시간과 충돌 회피에 초점을 맞추어 학습하여 탐색 증가로 인해 학습 중 도로를 이탈하는 시도를 더 자주했기 때문인 것으로 추측할 수 있다. 이 문제를 해결하기 위해서는 학습 시 1단계에서 안정화 조건에 도로 이탈 횟수를 추가하여 특정 임계값 미만일 때에만 2단계로 진행하도록 Teacher-Student Framework의 기준을 조정하는 것이 필요할 것이다.

5. 결론

본 논문은 복잡한 도로 주행 환경인 highway-env의 racetrack-v0에서 심층 강화학습 에이전트의 학습 효율성과 최종 성능을 향상시키기 위해 자동화된 CL 방법을 제안하고 그 효과를 실험을 통해 검증하였다. 제안하는 CL 방법은 시뮬레이션 환경의 복잡도를 결정하는 핵심 변수인 상대 차량의 최대 수를 자동으로 점진적으로 늘려감에 따라 학습 초반에는 기본적으로 도로를 따라 주행하는 방법을 학습하고, 학습 후반에는 주위 자동차와 충돌하지 않으면서 주행하는 방법을 학습한다. 이때 cooldown 기법과 interleaving 기법을 적용하여 학습의 안정성을 높혔다.

이 방법을 CL을 적용하지 않은 기준 모델과 비교한 결과, 제안된 방법이 학습 과정에서 훨씬 더 안정적으로 높은 보상과 긴 에피소드 길이에 수렴하는 것을 보였다. 또한 최종 성능 평가에서 제안하는 방법이 보상과, 스텝 수, 에피소드 길이, 충돌 횟수 측면에서 모두 통계적으로 유의미하게 우수한 결과를 달성하였다. 이는 제안하는 방법이 초기에 기본적인 자동차 주행 방법을 학습한 뒤, 복잡한 환경에서 회피하는 전략까지 순차적으로 학습하도록 유도하여 지역해 문제를 완화하고 더 우수한 정책을 획득하게 했음을 보여준다.

향후 연구로는 본 연구에서 학습된 정책이 한 번도 보지 못한 새로운 주행 환경에서 얼마나 잘 작동하는지 일반화 성능을 평가하는 것이다. 또한 제안하는 방법을 통해 새로운 주행 환경을 학습할 때에도 여전히 우수한 성능을 유지하는지 검증하는 것도 의미가 있을 것이다.

References

[1] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, "Deep Reinforcement Learning for Autonomous Driving: A Survey", IEEE Transactions on Intelligent Transportation Systems, vol. 23, iss. 6, June 2022, pp. 4909-4926, doi: 10.1109/TITS.2021.3054625.

[2] P. Ladosz, L. Weng, M. Kim, H. Oh, "Exploration in deep reinforcement learning: A survey", Information Fusion, vol 85, September 2022, pp. 1-22, doi: 10.1016/j.inffus.2022.03.003.

[3] Y. Bengio, J. Louradour, R. Collobert, J. Weston, "Curriculum learning", In Proceedings of the 26th Annual International Conference on Machine Learning, June 14-18, 2009, Montreal Quebec, Canada, pp. 41-48, doi: 10.1145/1553374.1553380.

[4] D. Lee, "Comparison of Reinforcement Learning Activation Functions to Maximize Rewards in Autonomous Highway Driving", The Journal of the Institute of Internet, Broadcasting and Communication, vol. 22, no. 5, October 2022, pp. 63-68, doi: 10.7236/JIIBC.2022.22.5.63.

[5] M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. Wiele, V. Mnih, N. Heess, J. T. Springenberg, "Learning by Playing Solving Sparse Reward Tasks from Scratch", Proceedings of the 35th International Conference on Machine Learning, vol. 80, February 2018, pp. 4344-4353, doi: 10.48550/arXiv.1802.10567.

[6] L. Anzalone, P. Barra, S. Barra, A. Castiglione, M. Nappi, "An End-to-End Curriculum Learning Approach for Autonomous Driving Scenarios", in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 10, October 2022, pp. 19817-19826, doi: 10.1109/TITS.2022.3160673.

[7] A. Graves, M. G. Bellemare, J. Menick, R. Munos, K. Kavukcuoglu, "Automated Curriculum Learning for Neural Networks", Proceedings of the 34th International Conference on Machine Learning, vol. 70, August 2017, pp. 1311-1320, doi: 10.48550/arXiv.1704.03003

[8] C. Florensa, D. Held, X. Geng, P. Abbeel, "Automatic Goal Generation for Reinforcement Learning Agents", Proceedings of the 35th International Conference on Machine Learning, vol. 80, July 2018, pp. 1515-1528, doi: 10.48550/arXiv.1705.06366.

[9] R. Campbell, J. Yoon, "Automatic curriculum learning with gradient reward signals", arXiv preprint, December 2023, pp. 1-11, doi: 10.48550/arXiv.2312.13565.

[10] D. Lee, "Comparison of Reinforcement Learning Algorithms for a 2D Racing Game Learning Agent", The journal of the institute of internet, broadcasting and communication, vol. 20, no. 1, February 2020, pp. 171-176, doi: 10.7236/JIIBC.2020.20.1.171.

[11] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, "Overcoming catastrophic forgetting in neural networks", Proc. Natl. Acad. Sci. U.S.A., vol. 114, no 13, March 2017, pp. 3521-3526, doi: 10.1073/pnas.1611835114.