자율운항선박의 안전운항을 위한 적대적 AI 및 강화학습 기술의 연구

홍예령[1], 박주현[2], 조혜원[3], 김민솔[4], 한지은[5], 이규영[6]
[1]성신여자대학교 AI융합학부
[2]동덕여자대학교 데이터사이언스전공
[3]이화여자대학교 전자전기공학과
[4]홍익대학교 컴퓨터공학과
[5]강남대학교 데이터사이언스학과
[6]한국과학기술원 정보보호대학원 박사수료
iamyerung@gmail.com, hhyyn00@gmail.com, jhw5974@gmail.com,
minsolkim0920@gmail.com, gks3177@naver.com, leeahn1223@kaist.ac.kr

A Study on Adversarial AI and Reinforcement Learning Technologies for Safe Navigation of Autonomous Ships

Ye-Ryeong Hong[1], Ju-Hyun Park[2], Hye-Won Jo[3],
Min-Sol Kim[4], Ji-Eun Han[5], Gyu-Young Lee[6]
[1]School of AI Convergence, Sungshin Women's University
[2]Dept. of Data Science, Dongduk Women's University
[3]Dept. of Electronic & Electrical Engineering, Ewha Woman's University
[4]Dept. of Computer Engineering, Hongik University
[5]Dept. of Data Science, Kangnam University
[6]Graduate School of Information Security, KAIST

요 약

해운산업에 인공지능 기술이 결합됨에 따라 자율 항해 기술이 급격히 발전하고 있다. 선박 운항의 안정성과 효율성을 높이기 위해 강화학습을 이용한 충돌 방지 및 경로 생성 연구가 활발히 이루어지고 있으나, 인공지능은 적대적 공격에 취약하다는 한계점이 있다. 이에 본 논문에서는 선박의 안전 운항을 위협하는 적대적 공격기법을 비교 분석하고, 강화학습 기술을 평가하여 가장 적합한 기법을 제안함으로써, 향후 선박 운항을 위한 연구 방향을 제시하고자 한다.

1. 서론

자율운항선박은 사람의 개입 없이 자동화된 시스템을 활용하여 운항하는 선박이며 인공지능과의 결합을 통해 급격히 발전하고 있다. 이는 빠르게 적용되고 있으며 해양분야에서의 안전성과 신뢰성 및 효율성을 크게 증가시킬 것으로 기대된다[1].

본 논문에서는 AI기술을 탑재한 자율운항선박의 안전한 운항을 위하여, 강화학습 기술을 구현하여 비교하고, 인간과 동물이 경험을 습득하여 지능을 형성해 가는 과정을 모방한 강화학습 Agent를 적용하여 해상 장애물과의 충돌을 방지하기 위한 연구 방안을 제시하고자 한다.

2. 관련 연구

2-1. 적대적 공격(Adversarial attack) 기술

원본 이미지에 의도적으로 섭동을 추가하여 딥러닝 모델의 오분류를 의도하는 공격 기술이며, 주요한 2가지 기술은 아래와 같다.

(1) FGSM (Fast Gradient Sign Method)

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x L(C(x,w),y)) \quad (1)$$

손실 함수(L)에 대한 입력 이미지(x)의 그라디언트를 계산한 후, 그 값이 커지는 방향으로 ϵ만큼의 잡음을 더하여 적대적 이미지를 생성한다.

(2) JSMA (Jacobian-based Saliency Map Attack)

$$\nabla F(X) = \begin{bmatrix} \frac{\partial Y_1}{\partial X_1} & \frac{\partial Y_1}{\partial X_2} & \cdots & \frac{\partial Y_1}{\partial X_n} \\ \frac{\partial Y_2}{\partial X_1} & \frac{\partial Y_2}{\partial X_2} & \cdots & \frac{\partial Y_2}{\partial X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_m}{\partial X_1} & \frac{\partial Y_m}{\partial X_2} & \cdots & \frac{\partial Y_m}{\partial X_n} \end{bmatrix}$$

그림1 Jacobian Matrix

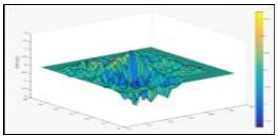


그림2 Saliency Map

그림(1)의 자코비안 행렬은 신경망의 출력에 대한 입력의 편미분들로 이루어진 행렬이며 입력 이미지의 각 픽셀이 출력에 미치는 영향을 계산한다.

그림(2)의 살리언시 맵은 행렬의 계산 결과인 각 픽셀의 영향을 정량적으로 나타내는 시각적 지도이며, 최종적으로 영향이 큰 픽셀 2개를 적대적 이미지를 생성에 반영한다.

2-2. 강화학습(Reinforcement Learning) 기술

강화학습은 에이전트가 현재의 상태(state)에서 특정 행동(action)을 취하고, 누적 보상(reward)을 최대화하는 방식으로 학습하는 AI모델이다.

(1) Q-Network

SARSA의 계산 복잡도, 차원의 저주 문제를 해결하기 위해 인공신경망이 Q-함수를 근사하여 상태, 행동에 따른 Q-값을 예측한다.

(2) REINFORCE

$$\theta_{t+1} \sim \theta_t + \alpha \left[\nabla_\theta \log \pi_\theta(a|s) G_t \right] \quad (2)$$

기존의 가치 기반(value-based) 알고리즘과 달리, 인공신경망이 정책을 근사하는 정책 기반(policy-based) 알고리즘으로, 각 행동을 할 확률을 예측하여 연속적인 행동 공간에서도 활용 가능하다.

(3) 액터크리틱(Actor-Critic)

크리틱은(가치망)은 가치 함수를 추정하도록 학습하여 행동의 가치를 평가하고, 액터(정책망)는 현재 상태에서 행동을 선택하는 정책을 학습해 행동을 최적화하는 강화학습 기법이다. 대표 알고리즘으로 A2C(Advantage Actor-Critic), A3C(Asynchronous Advantage Actor-Critic) 등이 있다.

3. 실험

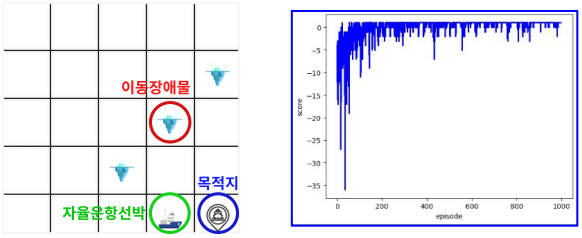
CIFAR-10 데이터셋으로 모델 학습 후, 공격기법별 적대적 이미지를 생성한 결과는 다음과 같다.

그림3 FGSM 이미지

그림4 JSMA 이미지

FGSM의 경우, 공격정확도(원본과 다른 클래스로 분류할 확률)는 94.029%로 높았으나, 원본과 적대적 이미지 간의 큰 차이가 육안으로 확인됐다. 반면 JSMA는 공격 정확도가 98.11%로 높게 기록되었고, 육안 상으로도 차이가 거의 없었다.

그림(5)-(a)에서 2차원 격자공간에 자율운항선박, 이동장애물, 목적지로 구성된 모의환경을 구축했다.



(a) 자율운항 모의환경 (b) 장애물 충돌 추이

그림5 Q-Network 자율운항 환경 및 성능 결과

좌표값 state, 상하좌우 4가지 action 그리고 목표도달 시 +1, 장애물 충돌 시 -1의 reward를 설정하였으며, 이동장애물이 있어 난이도가 높기 때문에 SARSA나 Q러닝이 아니라 딥러닝기술을 적용한 Q-Network 알고리즘을 사용하였다.

1000 에피소드로 학습을 진행한 결과, 그림(5)-(b)와 같이 이동장애물 충돌횟수가 점차 감소하여 수렴하는 것을 볼 수 있다.

학습훈련을 완료한 후 100 에피소드로 성능테스트를 진행한 결과, 이동장애물에 한번도 충돌하지 않았으며, 이로써 매우 높은 성능의 에이전트를 성공적으로 만들어 낸 것을 확인하였다.

4. 결론

본 연구를 통해 JSMA가 원본 이미지의 왜곡을 최소화하면서도 인식모델의 정확도를 가장 효과적으로 낮추는 적대적 공격기술인 것을 확인하였고, 이동장애물과 충돌 없이 완벽히 운행한 Q-Network 에이전트의 실험결과는 강화학습이 선박의 자율운항을 실현하는 데 매우 적합한 기술임을 입증하였다.

향후 자율운항선박의 속도, 방향과 같은 연속적 의사결정까지 고려할 때 Policy Gradient 계열의 REINFORCE 알고리즘이 우수하며, 특히 가치기반과 정책기반 강화학습을 결합한 액터크리틱 강화학습 기법이 가장 적합할 것으로 판단된다.

ACKNOWLEDGEMENT

※ 본 논문은 해양수산부 실무형 해상물류 일자리 지원사업(스마트해상물류 x ICT멘토링)을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] 이광일, 황승욱, '표준·시험인증 기술동향 - 자율운항선박 국제표준화 동향', TTA 저널, 175, 115-122, 2018