적대적 AI 공격에 대한 해양선박 보안강화 연구

전준석 [1], 김희준 [2], 이현화 [3], 윤다영 [4], 박지우 [5], 이규영 [6]
[1] 서울과학기술대학교 산업공학과 ITM 전공 학부생
[2] 고려대학교 컴퓨터융합소프트웨어학과 학부생
[3] 서울과학기술대학교 산업공학과 산업정보시스템 전공 학부생
[4] 성신여자대학교 융합보안공학과 학부생
[5] 수원대학교 컴퓨터학부 컴퓨터 SW 전공 학부생
[6] 한국과학기술원 정보보호대학원 박사수료

jjsmodac221@seoultech.ac.kr, kktongkr@korea.ac.kr, pluto5338@gmail.com, dyydreamer@gmail.com,
jiwoo061@gmail.com, leeahn1223@kaist.ac.kr

A Study on Strengthening the Security of Marine Ships against Adversarial AI Attacks

Jun-Seok Jeon[1], Hee-June Kim[2], Hyun-Hwa Lee[3], Da-Young Yoon[4], Ji-Woo Park[5], Gyu-Young Lee[6]
[1, 3]Dept. of Industrial Engineering, Seoul National University of Science and Technology
[2]Dept. of Computer Convergence Software, Korea University
[4]Dept. of Convergence Security Engineering, Sungshin Women's University
[5]Dept. of Computer Software, University of Suwon
[6] Graduate School of Information Security, KAIST

요 약

본 연구에서는 해양 선박의 보안 강화를 위해, 다양한 적대적 공격에 대응할 수 있는 방어 전략을 제안한다. FGSM, BIM, PGD, DeepFool, JSMA 등 주요한 적대적 공격 기법을 구현하여 실험한 결과, 적대적 훈련을 적용한 모델이 공격 후에도 높은 정확도를 유지하는 것을 확인하였으며, 따라서 향후 자율운항을 위한 안전성 확보에 중요한 역할을 할 것으로 기대한다.

1. 서론

4 차산업의 발전으로 AI 기술이 산업 전반에 확산되면서, 해양산업에서 자율운항선박 시스템이 새로운 패러다임으로 자리 잡고 있다. 특히 2025 년까지 자율운항선박 기술 개발 및 상용화 기반을 마련하고 있으며, 해양산업의 경제적 시장규모는 약 170 조까지 확대될 것으로 전망된다[1].

그러나 적대적 공격, 데이터 조작 등 AI 시스템에 대한 여러 보안위협은 해양산업의 안정성을 심각하게 약화시킬 수 있다. 자율운항 선박으로 얻을 수 있는 경제적 이익이 큰 만큼, 이러한 공격에 노출된다면 그 피해 역시 막대할 수 있다.

본 연구에서는 5 가지의 적대적 공격기법을 구현하고, 이를 활용하여 다양한 공격 · 방어 실험을 수행하였다. 이를 통해 적대적 훈련 기술이 해양선박 시스템의 보안성을 강화하는 데 유의미한 기여를 할 수 있음을 입증하였다.

2. 관련 연구

적대적 공격은 그림 1 과 같이 원본 이미지에 인간이 인지할 수 없을 정도의 미세한 노이즈를 추가하여 AI 모델이 잘못된 예측을 하도록 유도하는 기술이며, 대표적인 주요 기법들에 대한 설명은 아래와 같다.

1) **FGSM** (Fast Gradient Sign Method)
입력 데이터에 대한 손실 함수의 gradient 를 이용해, 작은 크기의 노이즈(ε)를 더하여 모델이 잘못 분류하도록 유도하는 공격이다.

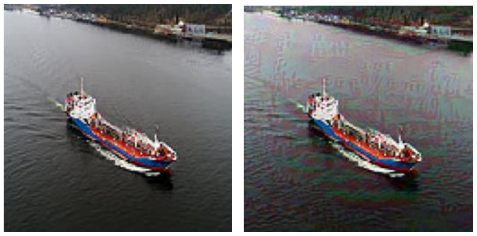
2) **BIM** (Basic Iterative Method)
FGSM기반의 공격으로, 여러 번의 공격을 통해 작은 노이즈를 반복하여 더욱 정교한 적대적 예제를 생성하는 기법이다.

3) **PGD** (Projected Gradient Descent)
BIM을 확장한 공격으로 원본 이미지에서 바로 공격을 시작하는 게 아니라 ε 범위 내에서

랜덤하게 선택된 지점에서 공격한다[2].

4) **DeepFool**
신경망 분류기의 결정 경계를 기반으로 한 반복적인 공격 알고리즘이다. 이 방법은 분류기의 결정 경계를 최소한으로 넘는 적대적 샘플을 생성하기 위해, 반복적으로 결정 경계를 추정하고 입력에 최소한의 교란을 가한다[3].

5) **JSMA** (Jacobian-based Saliency Map Attack)
Jacobian matrix을 통해 만들어진 saliency map을 활용한 적대적 공격 방법이다. 이미지를 분석하여 saliency map을 생성하고, 생성된 saliency map을 통해 수정할 픽셀을 결정한다. 모델이 타겟 클래스로 입력을 잘못 분류할 때까지 위 과정을 반복한다[4].



(그림 1) (a) 원본 이미지 (b) PGD 적대적 이미지

적대적 방어는 딥러닝 모델이 적대적 공격에 대해 견고한 성능을 유지하도록 하는 기법이다. 이는 모델이 교란된 데이터에 대해서도 올바른 예측을 할 수 있도록 훈련 과정에 다양한 방어 기법을 도입하는 것을 목표로 한다.

본 연구에서는 적대적 훈련(Adversarial Training)을 방어 기법으로 활용하였다. 적대적 훈련은 적대적 예제를 학습데이터에 추가하여, 적대적 이미지에 의한 공격에 AI모델을 강인하게 만들어 주는 것이다.

3. 실험

3.1 실험환경

본 연구에서는 Kaggle의 '**Game of Deep Learning: Ship Datasets**'을 활용하여, 5개 클래스(화물선, 군함, 항공모함, 크루즈선, 유조선)의 6,252개 훈련 이미지와 2,680개의 테스트 이미지를 사용했다.

훈련 이미지는 데이터 증강 기법 (랜덤 수평 뒤집기, -20° 회전, 1.1배 확대/축소)을 적용하였으며, Google Colab Pro 의 T4 GPU 환경에서 TensorFlow 2.17.0 를 사용하여 진행하였다.

선박 이미지 분류를 위해 **CNN** 모델을 활용하고, Adam optimizer(학습률:0.001)와 Sparse Categorical Crossentropy 손실함수를 사용하여 50 epoch 동안 학습하였으며, 과적합 방지를 위해 Dropout 과 Early Stopping 을 적용하였다.

3.2 실험결과

테스트 데이터셋에 대한 모델의 분류 정확도는 84.26%로 확인됐다. FGSM, BIM, PGD, DeepFool, JSMA 공격을 수행한 결과 AI모델의 분류성능이 크게 저하되는 현상을 관찰하였으나, 적대적 훈련을 적용한 이후 모델이 방어성능을 회복하는 모습을 확인하였다.

표 1 을 보면, PGD 공격 후 분류정확도가 0.1312 까지 감소하여, PGD 가 가장 강력한 공격력을 보였으나, 적대적 훈련 이후 PGD 공격에도 0.9168 의 높은 정확도를 유지하여 탁월한 방어효과를 확인할 수 있었다.

<표 1> 공격기법 별 적대적 훈련 전/후 성능 비교

Method	CNN 모델의 분류 정확도	
	적대적 훈련 전	적대적 훈련 후
FGSM	0.1855	0.7818
BIM	0.2137	0.6679
PGD	**0.1312**	**0.9168**
DeepFool	0.2398	0.8407
JSMA	0.2400	0.8217

4. 결론

본 연구에서는 해양선박에 대한 적대적 공격에 대응하기 위하여, 적대적 훈련을 방어 전략으로 선정하고 그 효과성을 검증하였다. 실험 결과 적대적 훈련 적용 시, 다양한 공격에서 높은 분류정확도를 유지하였고, 따라서 적대적 훈련이 해양 선박의 보안성을 강화하는데 기여할 수 있음을 입증하였다.

ACKNOWLEDGEMENT

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

참고문헌

[1] 박혜리, 박한선, 김보람, "자율운항선박 도입 관련 대응정책 방향 연구", 한국해양수산개발원, 2018.

[2] Madry, A, et al. "Towards deep learning models resistant to adversarial attacks", ICLR, 2018

[3] Moosavi-Dezfooli, et al. "DeepFool: a simple and accurate method to fool deep neural networks", CVPR, 2016.

[4] Combey et al. "Probabilistic Jacobian-based Saliency Maps Attacks", Machine Learning and Knowledge Extraction, 2020.