자율운항선박 보호를 위한 적대적 AI 기술의 연구

김수영[1], 박성진[2], 손건희[3], 이지민[4], 이규영[5]
[1]성신여자대학교 융합보안공학과 학부생
[2]수원대학교 정보보호학과 학부생
[3]국립공주대학교 인공지능학부 학부생
[4]성신여자대학교 컴퓨터공학과 학부생
[5]한국과학기술원 정보보호대학원 박사수료
suyeongg092@gmail.com, psj04100410262635@gmail.com,
songunhee5426@gmail.com, jm1004_lee@naver.com, leeahn1223@kaist.ac.kr

A Study of Adversarial AI techniques for Protecting Autonomous Ships

Su-Yeong Kim[1], Sung-Jin Park[2], Kun-Hee Son[3], Ji-Min Lee[4], Gyu-Young Lee[5]
[1]Dept. of Convergence Security Engineering, Sung-shin Women's University
[2]Dept. of Information Protection, Su-won University
[3]Dept. of Artificial Intelligence, Kong-ju National University
[4]Dept. of Computer Science and Engineering, Sung-shin Women's University
[5]Graduate School of Information Security, KAIST

요 약

자율운항 선박에 대한 기술 연구가 많은 관심을 받고 있지만, 그 근간을 이루는 인공지능 기술은 보안공격에 매우 취약하다. 본 논문에서는 선박이미지를 학습한 CNN AI 모델에 FGSM 및 BIM 공격을 가한 후 그 영향도를 비교하여 분석하였다. 그 결과 Adversarial Training 방어기법이 적대적 AI 공격을 효율적으로 차단할 수 있음을 실험을 통해 입증하였다.

1. 서론

자율운항선박은 인공지능, 사물인터넷, 빅데이터, 센서 등 모든 디지털 핵심 기술을 융합해 스스로 최적항로를 설정하고 항해할 수 있는 차세대 고부가가치 선박이다[1].

AI 기술은 자율운항 선박의 근간이 되는 기술이나, AI 보안 위협으로 인해 심각한 인명피해, 재산 피해를 초래할 수 있어 각종 위협과 공격에 대한 대비가 매우 필요한 실정이다.

이미지에 미세한 교란을 추가해 학습모델의 예측을 왜곡하는 적대적 공격을 차단하기 위해, 본 논문에서는 Adversarial Training 방어기법을 제안한다. FGSM, BIM 공격 전후의 정확도와 손실도를 실험을 통해 산출 및 분석하여 그 유용성을 입증하였다.

2. 관련 연구

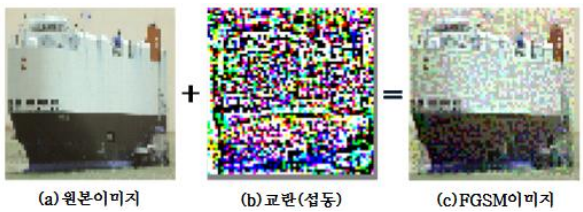
적대적 예제(Adversarial Example)를 생성하는 주요 기술에 대한 설명은 다음과 같다.

2.1 FGSM (Fast Gradient Sign Method)

FGSM은 적대적 예제를 생성하는 단순하고 일회성인 Non-Targeted 공격기법이다. 이 기법은 모델의 예측값과 정답 간의 손실함수 오차를 계산하고, 이 오차를 입력값으로 편미분한 후, 해당 오차를 최대화하는 기울기 방향으로 고정 크기의 교란(perturbation)을 원본 이미지에 1회 추가한다.

$$x_{adv} = x + \varepsilon \cdot sign(\nabla_x L(C(x,w),y)) \quad (1)[2]$$

식(1)은 FGSM 생성 수식이고, 그림(1)은 원본이미지와 ϵ=0.1 옵션으로 생성한 FGSM 이미지이다.



(그림 1) FGSM 적대적 이미지 생성결과

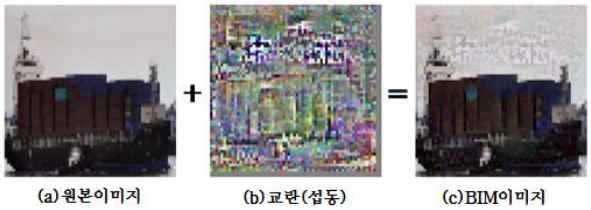
2.2 BIM (Basic Iterative Method)

FGSM을 보다 개선한 기술이며, 작은 규모의 섭동을 반복적으로 적용한다.

$$x_{n+1}^{adv} = Clip\left\{x_n^{adv} + a \cdot sign\left(\nabla_{x_n^{adv}} J\left(x_n^{adv},\ y\right)\right)\right\} \quad (2)[3]$$

식(2)은 BIM 생성 수식이고, 그림(2)은 원본이미지

와 α=0.01 옵션으로 생성한 BIM 이미지이다.



(그림 2) BIM 적대적 이미지 생성결과

3. 실험

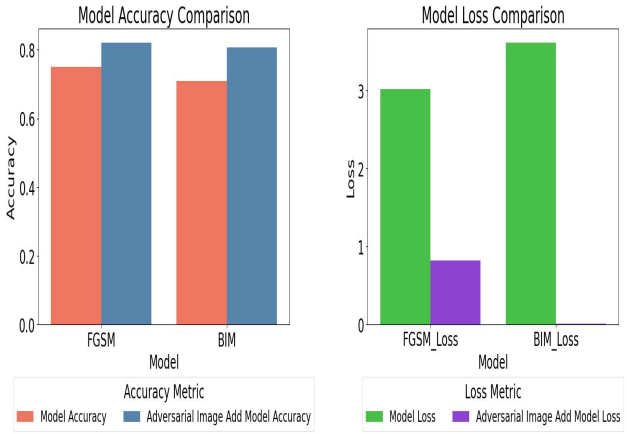
3.1 실험 환경

데이터셋은 kaggle에서 제공하는 ship image dataset를 사용하였고, 3계층 2D CNN 모델을 구축하고 20 epoch, adam 옵티마이저, 32 배치로 설정하여 학습을 진행하였다.

상기 학습과정을 통해 baseline model을 만들고, 이후 FGSM, BIM 기법으로 제작한 적대적 이미지로 공격을 진행하여 정확도와 손실률을 측정하였다.

3.2 실험 결과

훈련 데이터셋은 8000개를 이용해서 훈련을 시키고 이후 테스트셋에 평가를 진행했다. 그림(3)과 같이 FGSM과 BIM 공격 이후 성능을 확인 할 수 있다.



(그림 3) 적대적 훈련 전 후 성능 비교(좌측은 정확도, 우측은 loss)

BIM이 FGSM보다 낮은 성능을 보이는 것을 확인할 수 있었고, 해당 이유로는 여러번의 작은 변화를 통해서 더 강력한 적대적 예제 만들어서 모델의 정확도 낮게 나온 것을 확인할 수 있었다.

이후 추가적으로 모델의 강건성을 위해서 적대적 예제를 훈련세트에 추가하여 학습을 진행하였다.

본 논문에서는 각 공격방식이 적용된 이미지 4000개씩 추가하여 총 16000개의 훈련세트를 구성했다.

<표 1> 적대적 공격 및 방어 성능 비교

성능 항목	적대적훈련 前		적대적훈련 後	
	FGSM	BIM	FGSM	BIM
정확도	0.74	0.70	0.82	0.80
손실	3.02	3.62	0.82	0.01

표(1)에서 훈련 후 모델의 정확도가 약 70% 정도 상승한 것을 확인할 수 있다. 성능향상을 가져온 이유는 노이즈 데이터를 추가로 학습해서 일반화 성능이 높아진 것으로 예상된다.

4. 결론

본 논문에서는 2D CNN 기반의 선박 이미지 분류 모델을 대상으로 적대적 훈련을 적용하여 적대적 공격에 대한 방어 성능을 크게 향상시켰다. 훈련 후 적대적 예제에 대한 모델의 정확도는 크게 약 70% 정도 개선되었으며, 특히 BIM 공격에 대한 저항력이 강화되었다.

이러한 실험을 통해 Adversarial Training이 적대적 공격으로부터 자율운항선박을 보호하기 위해 필수적임을 확인하였다.

향후에는 다양한 적대적 공격 기법에 대해 더 강력한 방어 메커니즘을 연구할 예정이다.

ACKNOWLEDGEMENT

- 본 논문에 참여한 저자들은 모두 공동1저자이며, 논문작성에 기여한 정도가 같습니다.
- 본 논문은 해양수산부 실무형 해상물류 일자리 지원사업(스마트해상물류 x ICT멘토링)을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] 유지운, “자율운항 선박의 인공지능 : 잠재적 사이버 위협과 보안”, 고려대학교 정보보호대학원, 고려대학교 정보보호대학원 학위논문(석사), p.3, 2023.

[2] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, "Explaining and Harnessing Adversarial Examples", International Conference on Learning Representations (ICLR), Banff, 2015, pp. 1-11.

[3] Chen, Z., Luo, W., Naseem, M. L., Kong, L., & Yang, X. Comprehensive comparisons of gradient-based multi-label adversarial attacks, Complex & Intelligent Systems, 2024, pp. 1-15.