# 자율주행 선박의 적대적 공격에 대한 신경망 모델의 성능 비교

허태훈[1],김주형[2],김나현[3],김소연[4]

[1]한림대학교 소프트웨어학부 빅데이터학과
[2]한국공학대학교 전자공학부 전자공학전공
[3]중앙대학교 국제물류학과
[4]서경대학교 소프트웨어학과

taehoon121@hallym.ac.kr, ju99119@tukorea.ac.kr, nhk2903@cau.ac.kr, qws1566@skuniv.ac.kr

# Performance Comparison of Neural Network Models for Adversarial Attacks by Autonomous Ships

Tae-Hoon Her[1], ,Ju-Hyeong Kim[2], Na-Hyun Kim[3],So-Yeon Kim[4],
[1]Dept. of Bigdata, Hallym University
[2]Dept. of Electronics, Tech University of Korea
[3]Dept. of International Logistics Chungang University
[4]Dept. of software, Seokeong University

**요 약**

자율주행 선박의 기술 발전에 따라 적대적 공격에 대한 위험성이 대두되고 있다. 이를 해결하기 위해 본 연구는 다양한 신경망 모델을 활용하여 적대적 공격을 탐지하는 성능을 체계적으로 비교, 분석하였다. CNN, GRU, LSTM, VGG16 모델을 사용하여 실험을 진행하였고, 이 중 VGG16 모델이 가장 높은 탐지 성능을 보였다. 본 연구의 결과를 통해 자율주행 선박에 적용될 수 있는 보안 모델 구축에 대한 신뢰성 있는 방향성을 제시하고자 한다.

## 1. 서론

4 차 산업 혁명 시대가 도래함에 따라 인공지능 기술이 자율주행 선박에까지 적용되기 시작했다. 특히 2018 년 5 월 국제해사기구(IMO)가 해사 안전, 보안 관련 14 개 국제 협약 제정 착수에 합의한 이후로부터 국제적으로 자율주행 선박 기술 개발과 시험 항해가 활성화되었다. [1]

한편 자율주행 선박이 활성화되기까지 아직 해결되지 않은 일부 문제점들도 남아있는데, AI 보안 취약점을 분석하여 공격하는 적대적 공격의 위험에 대한 방어책이 아직 제대로 마련되어 있지 않다는 점이다. 특히 수출입 물동량의 99.7%가 해상 운송으로 이루어지는 대한민국 산업을 고려하면, 적대적 공격 탐지 및 방어에 대한 연구는 매우 중요하고 꾸준히 계속되어야 한다. [2]

따라서 본 연구에서는 VGG16 등 신경망 모델을 사용하여 적대적 공격을 탐지하고 그 성능을 비교 분석한 후 자율주행 선박에 대한 딥러닝 보안 모델 구축의 방향성을 제시하고자 한다.

## 2. 적대적 공격 유형 및 특징

적대적 공격은 딥러닝의 심층 신경망을 이용한 모델에 적대적 교란(Adversarial Pertubation)을 적용하여 오분류를 유발하고 신뢰도 감소를 야기하는 머신러닝 공격 기법이다. 적대적 공격의 종류는 공격대상이 되는 모델에 관련된 모든 정보를 알고 공격하는 화이트박스 공격과 일부 정보만 알고 공격하는 블랙박스 공격이 있다. 가장 대표적인 적대적 공격 방식으로는 FGSM 이 있다. FGSM 은 화이트 박스 공격 중 기울기를 활용하여 정상적인 이미지에 사람의 눈으로 인식할 수 없는 노이즈를 추가하여 모델이 잘못된 분류를 하게 만드는 적대적 공격이다.[3]

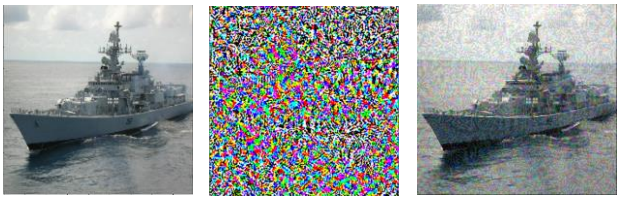
## 3. FGSM 기법

본 연구에서는 FGSM(Fast Gradient Signed Method) 기법을 이용하여 적대적 공격 데이터 셋을 제작하였고, 이를 방어 모델의 성능 확인에도 사용하였다. 원본 이미지에 특정한 작은 왜곡을 추가하여 신경망으

로 하여금 항공모함을 퍼즐이미지로 인식하도록 만들었다. 입력이미지 각 픽셀의 손실에 대한 기여도를 그래디언트를 통해 계산한 후, 그 기여도에 따라 픽셀값에 왜곡을 추가함으로써 생성할 수 있다.[4]

$$x + \epsilon \mathrm{sign}(\nabla_x J(\theta, x, y))$$

(식 1) [4]



(그림 1) 왼쪽:원본 이미지, 가운데: epsilon=0.150, 오른쪽: 생성된 적대적 이미지

## 4. 적대적 탐지 모델 성능 비교 분석

본 연구의 주 목적은 적대적 공격을 탐지하기 위한 다양한 신경망 모델의 성능을 체계적으로 비교 분석하는 것이다. 실험은 CNN, GRU, LSTM, VGG16 모델을 활용해서 이루어졌다. 각 모델은 동일한 컴퓨팅 환경에서 NVIDIA RTX 2080TI GPU 를 사용하여 학습을 진행하였다.

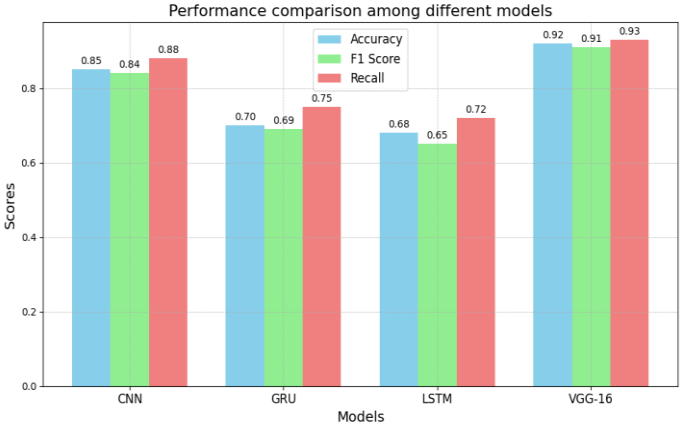
학습은 총 50 epoch 동안 수행되며, 손실 함수로는 Cross Entropy Loss 를 사용하였으며 Learning rate 는 0.001 로 설정, 해당 값은 Grid Search 를 통해 여러 후보 값 중에서 최적의 성능을 보이는 값으로 선택되었다.

데이터셋은 COCO 데이터셋을 활용하여 초기 학습을 진행하였다. 과적합을 방지하고 모델의 일반화 능력을 향상시키기 위해 다양한 데이터 증강 기법을 적용하였다. 이미지 회전은 -20 도에서 +20 도까지, 반전은 수평과 수직, 확대/축소는 0.8 배에서 1.2 배까지 적용하였다. 드롭아웃 비율은 0.4 로 설정하여 신경망 내 노드의 일부를 무작위로 무시하는 방식으로 과적합을 방지하였습니다.

## 5. 실험 및 평가

성능 평가는 자체 수집한 선박 데이터 7557 장에 적대적 공격을 적용하여 생성한 이미지를 테스트 데이터로 사용하였다. 해당 테스트 데이터는 실제 선박 운항 환경에서 발생할 수 있는 다양한 적대적 공격 시나리오를 반영하도록 설계하였다. 성능 평가 지표로는 정확도, 정밀도, 재현율, F1 score 를 사용하였다. [그림 2]에서 확인할 수 있듯이 VGG16 모델이 F1 score 0.93 및 정확도 92%로 가장 높은 성능을 보였다. 이에 반해, CNN 모델은 F1 스코어 0.84 와 정확도 85%를 기록하였으며, GRU 와 LSTM 모델은 각각 F1 스코어 0.69, 0.65 와 정확도 70%, 68%를 보였다. 이러한 성능 차이는 LSTM 과 GRU 가 주로 시계열이나 텍스트 데이터에 적합하고 이미지 처리에는 상대적으로 덜 특화되어 있기 때문으로 판단된다.

이러한 결과를 통해 본 연구는 실제 선박 운항에서 적대적 공격이 적용될 가능성을 고려하면, 이미지 데이터에 대한 적대적 공격 탐지에 있어 CNN 과 VGG16 모델이 상대적으로 높은 성능을 보임을 확인하였다. 해당 결과는 LSTM 과 GRU 가 일반적으로 시계열 데이터나 텍스트 데이터에 더 특화되어 있다는 기존의 연구 결과와도 일치한다. 따라서, 신경망 구조의 선택은 탐지하고자 하는 데이터의 유형과 특성에 따라 신중히 결정되어야 함을 강조하고자 한다.




(그림 2)

## 6. 결론

본 논문은 자율주행 선박에서의 적대적 공격 탐지를 다루고, 다양한 신경망 모델의 성능을 비교 분석하여 이미지 데이터에 대한 적대적 공격 탐지에 있어 CNN 과 VGG16 모델이 상대적으로 높은 성능을 보임을 확인하였다. 이러한 결과를 통해 신경망 모델을 선택할 때 데이터 유형과 특성을 고려해야 함을 강조하며, 자율주행 선박의 보안 강화를 위한 방향성을 제시한다.

※ 본 논문은 해양수산부 실무형 해상물류 일자리 지원사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

## 참고문헌

[1] 양창호.(2018). IMO 국제해사 정책동향.한국해양수산개발원

[2] 백병선. (2011). 미래 한국의 해상교통로 보호에 관한 연구. 국방정책연구, 27(1), 165-200.

[3] Jeon Ho-been, Comparison of Deep Learning-based Image Recognition Disruption of Adversarial Example Generation Methods, Kangwon univ.(2023),p.1

[4] Ian J. Goodfellow, Explaining And Harnessing Adversarial Examples, ICLR 2015, 11