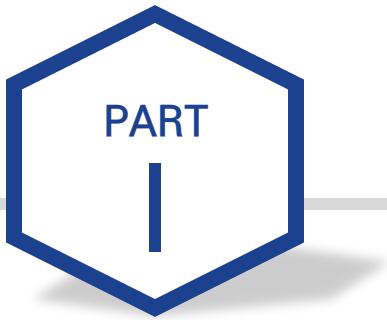
한국정보처리학회 ACK 2024 – KIPS_C2024B0147

적대적 AI 공격에 대한 해양선박 보안강화 연구

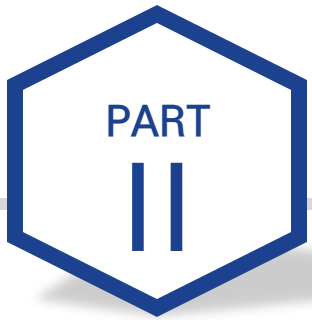
2024. 10.31

전준석, 김희준, 이현화, 윤다영, 박지우, 이규영



연구 배경 및 필요성

- 연구 개발의 배경
- 연구 최종 목표



관련 연구

- 적대적 공격
- 적대적 훈련

실험

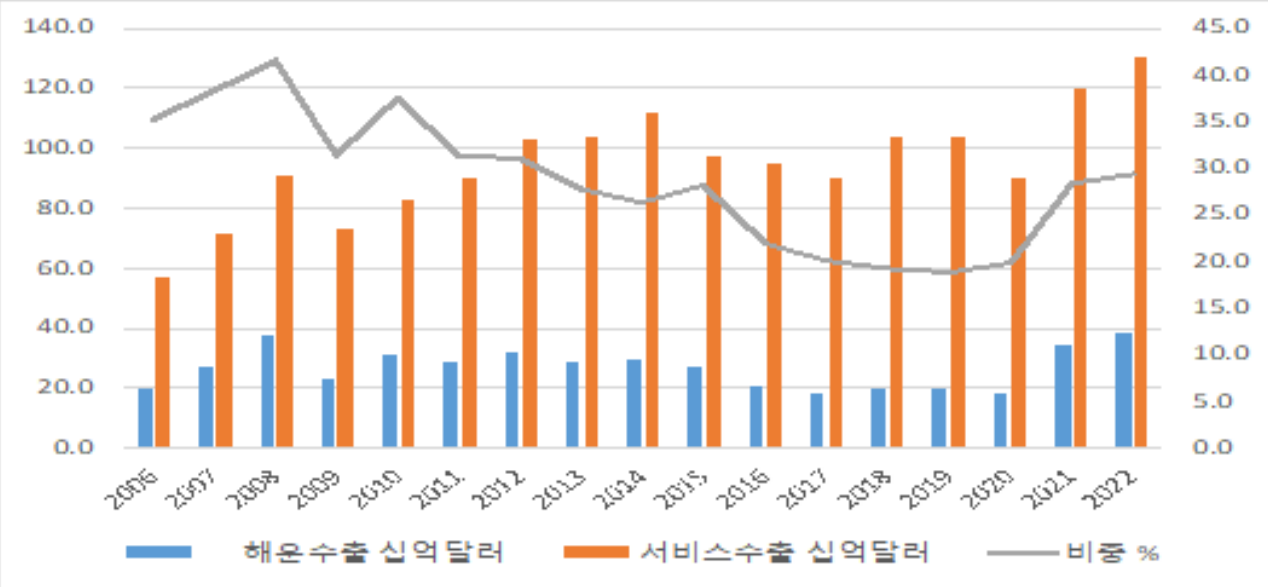
- 데이터셋
- 실험 프로세스

결론

- 결론

해운산업 및 자율운항선박의 발전

- ❖ 해운산업 사상 최대 수출실적 달성(2022년 기준)
  - 해운서비스 수출액 383억 달러, 한화 약 49.5조원 달성
- ❖ 2024년 상반기 선박 수주액 세계 1위 달성
- ❖ 대한민국, 선복량 세계 4위의 해운 강국
- ❖ 수출입 화물의 99.7%가 선박을 통해 운송되고 있음.(2020년 기준 수치)

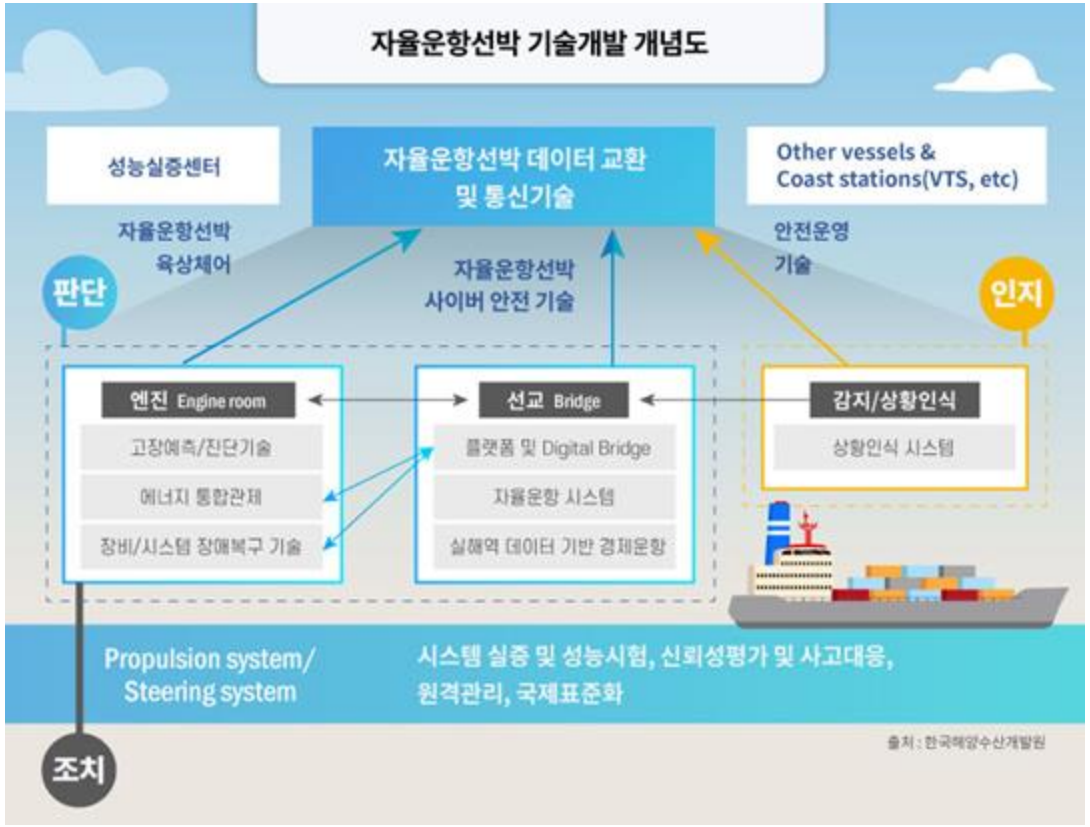


해운산업 및 자율운항선박의 발전

- ❖ 해운산업의 새로운 패러다임: 자율운항선박
  - 자율운항선박: AI, IoT, 빅데이터, 센서 등을 융합하여 선원의 의사결정을 지능화, 자율화된 시스템으로 대체할 수 있는 고부가가치의 선박
- ❖ 2025년 자율운항선박 상용화 목표, 글로벌 시장 규모 한화 약 170조까지 확대될 것으로 전망
- ❖ 전체 선박 시장 중에 자율운항선박이 차지하고 있는 규모가 50%를 넘을 것으로 예상

AI 시스템에 대한 보안 위협, 적대적 공격

- ❖ 자율운항선박으로 얻을 수 있는 경제적 이익 ↑, 그만큼 선박 시스템이 공격에 노출될 경우 피해가 막대할 수 있음.
- ❖ 대형 해양 사고의 우려 역시 존재



해양선박 분야 주요 사이버 보안 침해사고 사례

연도	분야·기업	분야	피해 사례
2017년	머스크	터미널 IT시스템	3주간 시스템 마비, 3000억 손실
	컨테이너선	선박 항해 시스템	10시간 동안 통제권 상실
2018년	바르셀로나 항만	항만 IT시스템	시스템 폐쇄 및 포렌식 의뢰
	샌디에이고 항만	항만 IT시스템	시스템 폐쇄 및 포렌식 의뢰
2019년	자동차 운반선	선박 IT시스템	대상 시스템 포맷
2021년	트랜스넷 SOC	항만 IT 시스템	모든 항만 터미널 운영 중단

(자료: 한국선급 기술정책제언연구집(임정규, 손금준·2020))

연구 목표

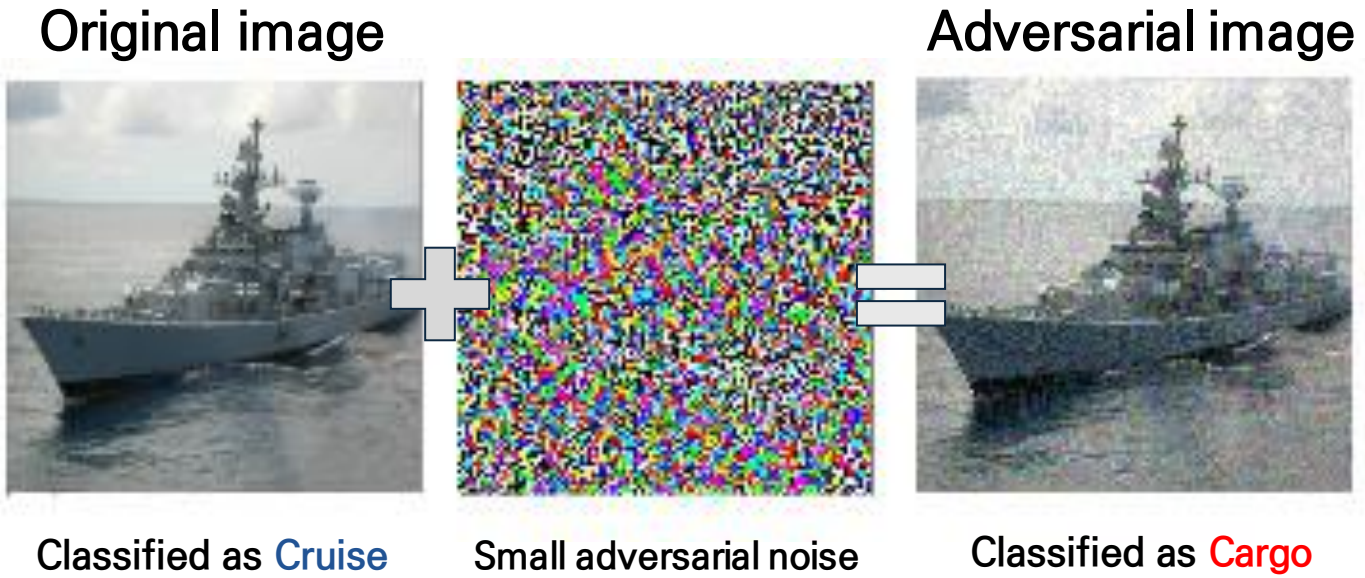
❖ **자율운항선박 시스템의 보안성을 강화하기 위해 적대적 AI 공격에 대응하는 효과적인 방어 기술을 개발하고 검증하는 것을 목표**

① **적대적 공격 기법 구현:** 자율운항선박 시스템을 대상으로 한 5가지 적대적 AI 공격 기법을 설계 및 구현
② **방어 기법 검증:** 다양한 공격 상황에서의 실험을 통해 기존 방어 기법의 효과를 분석
③ **해양선박 시스템의 보안성 평가:** 연구를 통해 제안된 방어 기법이 자율운항선박의 AI 시스템에 얼마나 유의미한 보안성을 제공하는지 실험적으로 검증

❖ **기대 효과**

자율운항선박 시스템의 **실질적인 보안 강화**에 기여하며, 적대적 AI 공격에 대한 **대응 전략 수립**에 중요한 참고자료 역할

적대적 공격 (Adversarial Attack)

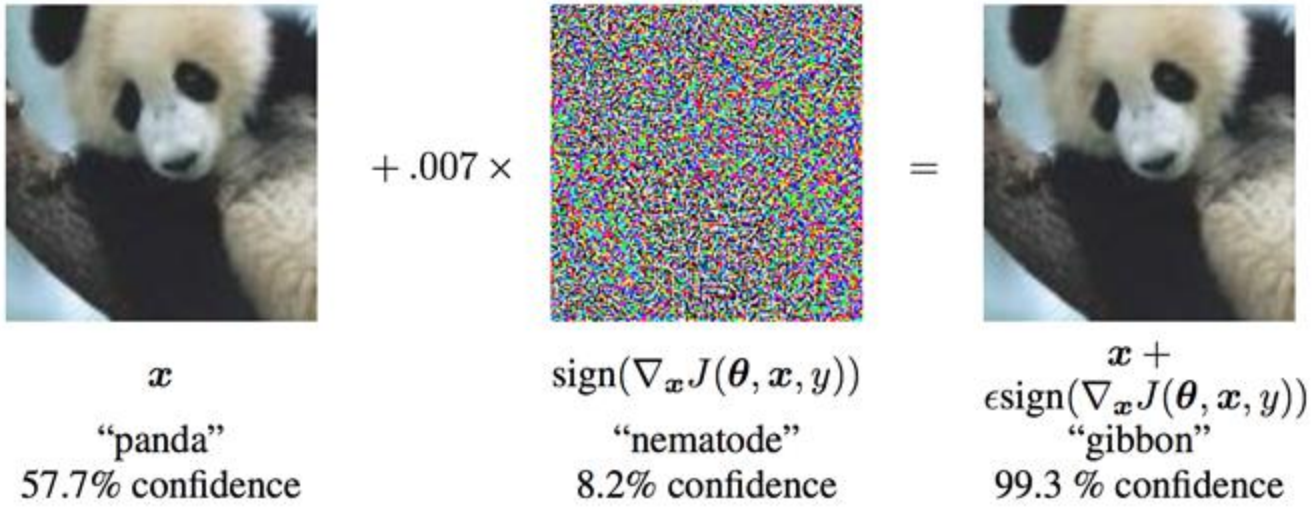


적대적 공격은 그림과 같이 원본 이미지에
인간이 인지할 수 없을 정도의 미세한 노이즈를 추가하여
AI모델이 잘못된 예측을 하도록 유도하는 기술

적대적 공격 – FGSM (Fast Gradient Sign Method)

: 딥러닝 모델의 취약성을 이용한 적대적 공격 기법으로, 입력 데이터에 작은 노이즈를 추가하여 모델이 잘못된 예측을 하도록 유도하는 방법

$$adv_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))$$



의의:

- 딥러닝 모델이 미세한 변화에도 취약함을 보여줌.
- 적대적 예시를 이용한 적대적 학습 등 방어 기법 개발 촉진.

한계:

- 단일 스텝 공격으로 방어 기법에 쉽게 대응 가능.
- I-FGSM, PGD와 같은 다중 스텝 기법이 이후에 제안됨.

적대적 공격 – BIM (Basic Iterative Method)

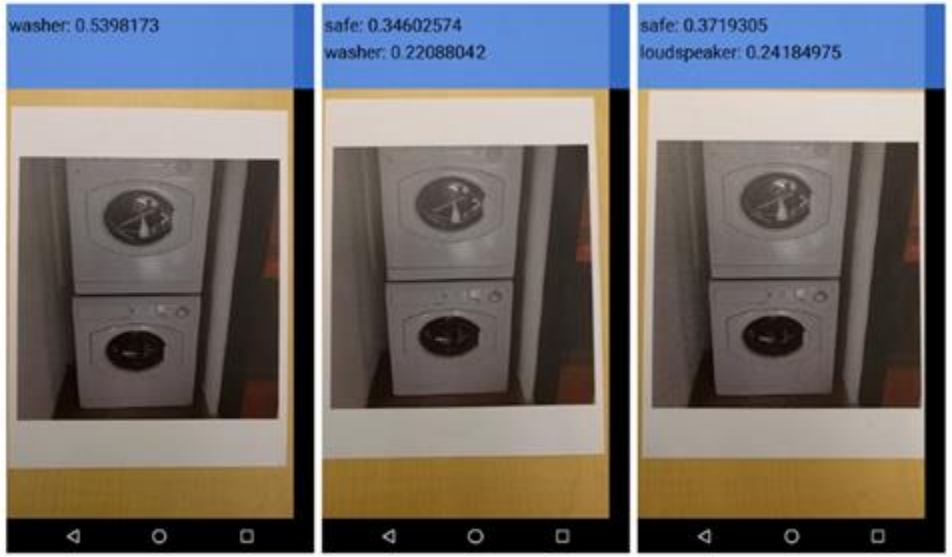
: FGSM기반의 공격으로, 여러 번의 공격을 통해 작은 노이즈를 반복하여 더욱 정교한 적대적 예제를 생성하는 기법

$$x_{adv}^{t} = x^{t-1} + \gamma \cdot \mathrm{sign}\left(\nabla_x \mathcal{L}(C(x^{t-1}, w), y)\right)$$

x – 원본 이미지(input 이미지)
t – 이터레이션(반복) 횟수를 의미
γ – 각 단계에서 추가되는 노이즈의 크기를 의미

- BIM 공격 예시(세탁기 이미지에 노이즈 추가)
  - (c) 적은 양의 노이즈(ϵ = 4/255)를 추가한 경우, 세탁기로 바르게 분류
  - (d) 많은 양의 노이즈(ϵ = 8/255)를 추가한 경우, 스피커로 잘못 분류

  → 노이즈 크기에 따라 이미지를 원래의 레이블대로 바르게 분류할 수도 있고, 잘못 분류할 수도 있음.



(b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

적대적 공격 – PGD (Projected Gradient Descent)

: FGSM기반의 BIM 공격을 더욱 발전시킨 방법으로, 원본 이미지에서 바로 공격을 시작하는 게 아니라 ε 범위 내에서 랜덤하게 선택된 지점에서 공격을 시작하는 방법

$$x_{adv}^{t} = \Pi_{\epsilon}\left(x^{t-1} + \gamma \cdot \mathrm{sign}\left(\nabla_x \mathcal{L}(C(x^{t-1}, w), y)\right)\right)$$

- $\Pi\epsilon$: ϵ–ball로의 투영 함수
- γ: 스텝 크기
- sign(·): 그래디언트의 부호
- ▽_x L : 손실 함수 $\mathscr{L}$의 입력 x에 대한 그래디언트
- C(x,w): 모델의 출력 (예측값)
- y: 원본 입력 데이터 x의 실제 레이블

원본 이미지

pgd 공격 이미지

pgd 공격을 이용해 선박 이미지에 적대적 공격을 한 결과

BIM 공격기법 차이

(−ϵ,ϵ) 범위 내에서 랜덤 노이즈를 더해 초기화 하고 이후 공격을 시작

FGSM	1번 공격
BIM	작은 노이즈를 반복하여 여러 번 공격
PGD	랜덤 노이즈를 더한 후, 여러 번 공격

적대적 공격 – JSMA (Jacobian-based Saliency Map Attack)

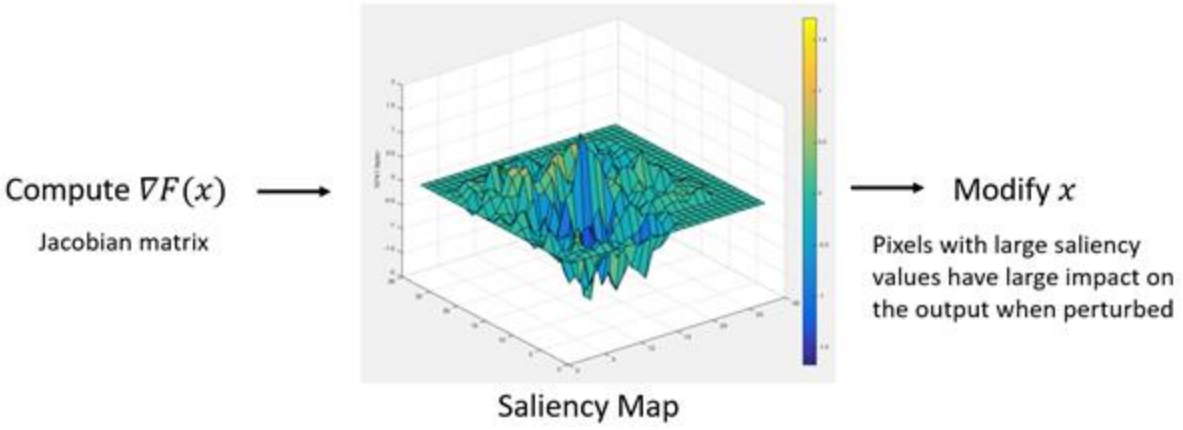
: Jacobian matrix을 통해 만들어진 saliency map을 활용한 적대적 공격 방법

* saliency map 정의 *

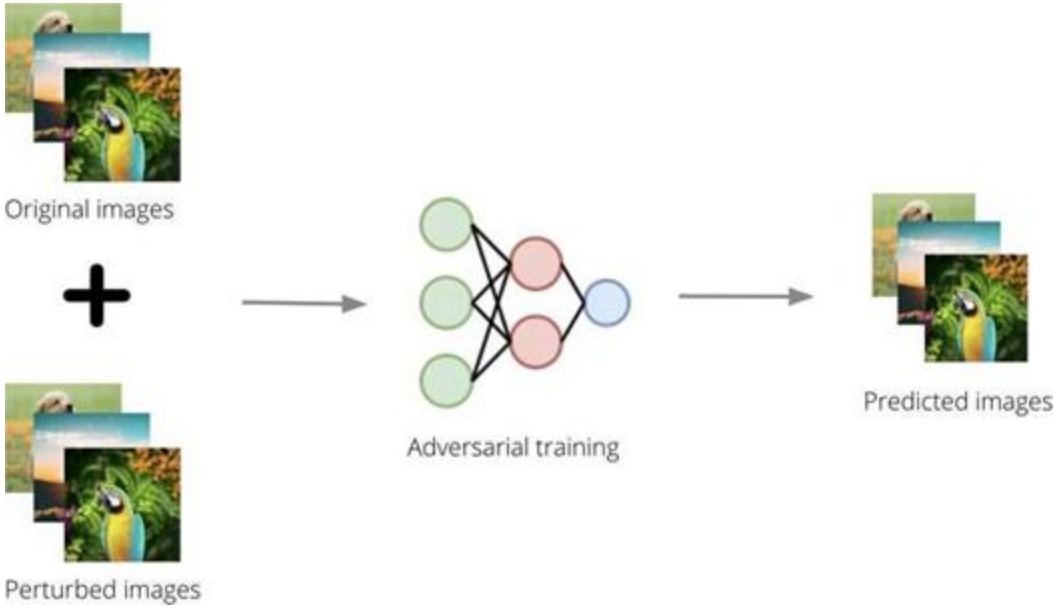
$$S[x,t][i] = \begin{cases} 0 & \text{if } \frac{\partial Z_t(x)}{\partial x_i} < 0 \text{ or } \sum_{k \neq t} \frac{\partial Z_k(x)}{\partial x_i} > 0 \\ \frac{\partial Z_t(x)}{\partial x_i} \cdot \left| \sum_{k \neq t} \frac{\partial Z_k(x)}{\partial x_i} \right| & \text{otherwise.} \end{cases}$$

- 입력 이미지의 각 픽셀이 출력에 미치는 영향을 Jacobian matrix로 계산
- saliency map을 통해 적대적 공격의 성공 확률을 증가시키는 픽셀을 찾아 수정

- white-box attack
- 가능한 적은 수의 픽셀만 변경해 잘못된 클래스를 예측하도록 유도
- 노름(L_0)을 기반으로 하여 픽셀 수를 최소화하면서도 모델을 효과적으로 속이는 것이 특징



적대적 방어 – Adversarial Training



➔ 모델이 적대적 예제를 학습 과정에 포함하여 훈련하는 방식.
적대적 학습은 모델이 원본 데이터뿐만 아니라 적대적 예제
(adversarial examples)에도 강건해질 수 있도록 설계.

구체적으로는, 모델이 훈련 중에 적대적 예시를 생성하고 이를 모델이 학습하도록 하여,
모델이 적대적 공격을 받았을 때에도 보다 높은 성능을 유지하도록 하는 방어 기술.

Sharma, Shorya. (2022). Multi-SAP Adversarial Defense for Deep Neural Networks. International Journal of Advanced Science Computing and Engineering. 4. 32-47. 10.30630/ijasce.4.1.76.

데이터셋

Cargo

Tanker

Military

Carrier

Cruise

- Kaggle의 'Game of Deep Learning: Ship Datasets' 사용
- 5개 클래스: 화물선, 군함, 항공모함, 크루즈선, 유조선 존재

➢ 훈련 이미지: 6,252개
➢ 테스트 이미지: 2,680개

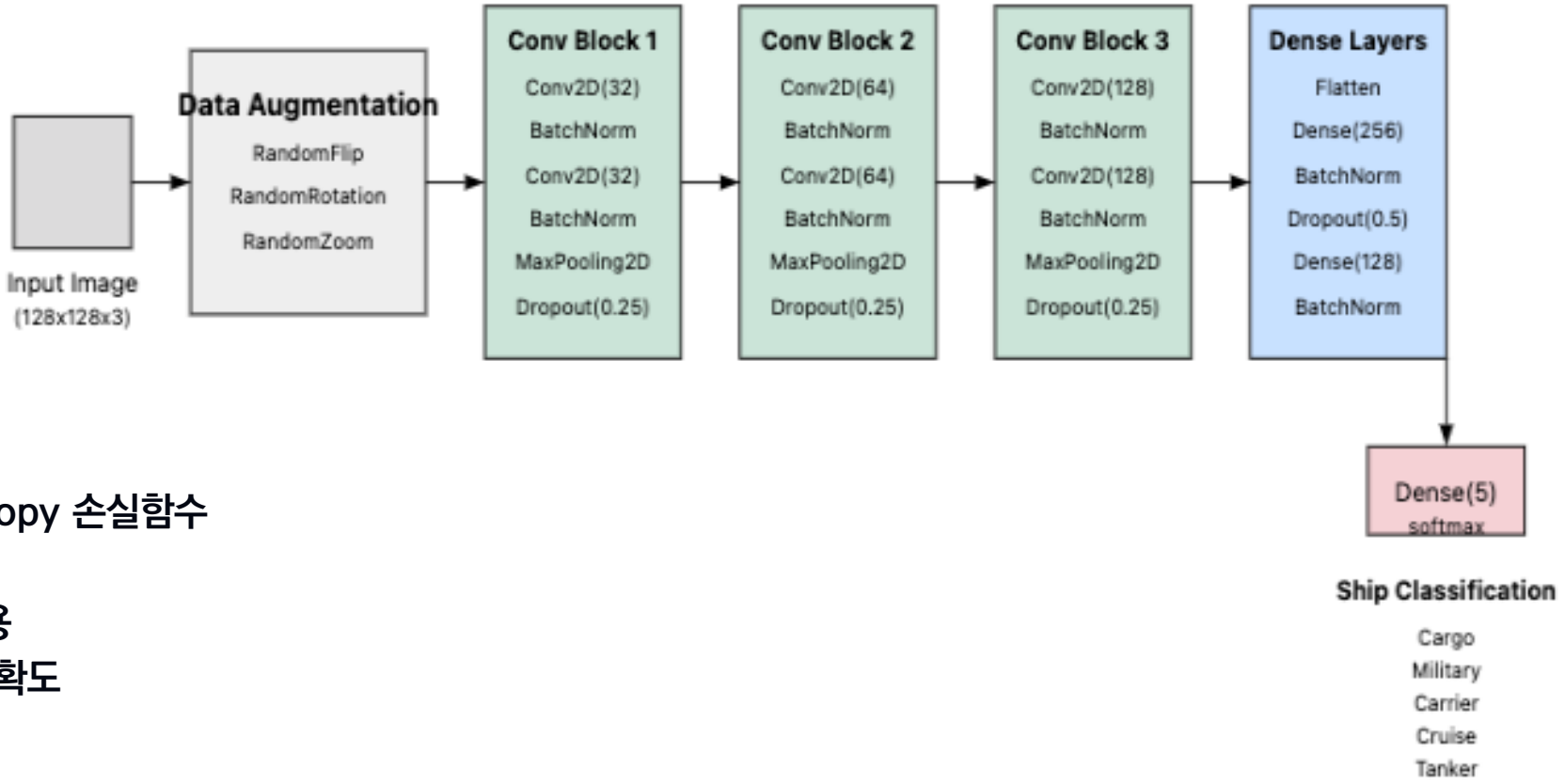
데이터 전처리

- 훈련 데이터셋 증강 기법 적용:
  –〉 랜덤 수평 뒤집기
  –〉 –20° 회전
  –〉 1.1배 확대/축소

실험 – CNN 모델 학습

실험 플랫폼

- Google Colab Pro의 T4 GPU 환경
- TensorFlow 2.17.0 사용



1. 모델 및 학습

- CNN 모델 사용
  - Adam optimizer (lr: 0.001)
  - Sparse Categorical Crossentropy 손실함수
  - 50 epoch 학습
  - DropOut과 Early Stopping 적용
  - 테스트 데이터셋에서 84.26% 정확도

실험 – 적대적 공격 및 적대적 훈련 수행

2. 적대적 공격 수행 (테스트 데이터셋)

- **테스트 데이터셋:** 2680개 이미지
- **적대적 공격 기법:** FGSM, BIM, PGD, DeepFool, JSMA 사용
- **결과:** 초기 분류 정확도에 비해 성능 저하 확인

3. 적대적 훈련 및 성능 개선

- **훈련 데이터셋:** 기존 훈련 데이터셋 + 적대적 공격을 받은 손상된 이미지 추가
- **적대적 훈련:** 모델이 공격에 대비할 수 있도록 재학습
- 학습하지 않은 손상된 데이터셋을 테스트셋으로 활용하여 각 공격별 결과 확인

실험 결과

적대적 공격	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

실험 결과

CNN 모델의 적대적 공격에 대한 분류 성능 평가

- **모델 정확도:** 기본 테스트셋에 대해 84.26%의 분류 정확도를 달성.
- **적대적 공격에 따른 성능 저하:** FGSM, BIM, PGD, DeepFool, JSMA공격에서 분류 성능이 크게 저하됨. PGD 공격 결과, 0.1312라는 분류 정확도로 가장 공격이 효과적이었음을 확인.
- **적대적 훈련 효과:** 적대적 훈련 후, 공격 방어 성능이 회복되었으며 특히 PGD 공격에서 0.9168까지 정확도가 향상.

결론

1. **적대적 훈련의 필요성:** 적대적 훈련은 해양 선박의 자율운항 시스템을 강화하기 위한 필수적인 방어 기술.

2. **적대적 공격에 대한 대응력:** 이 훈련을 통해 FGSM, PGD 등 다양한 적대적 공격에서도 모델이 높은 분류 성능을 유지할 수 있음을 확인.

3. **훈련 효과:** 적대적 훈련을 적용한 모델은 공격 후에도 최대 91% 이상의 정확도를 유지하여, 방어 효과가 뛰어남.

4. **AI 시스템의 신뢰성 향상:** 이로 인해 자율운항 선박의 보안성은 물론, AI 시스템의 신뢰성도 크게 향상될 수 있음을 증명.

5. **미래 적용과 연구의 필요성:** 적대적 훈련은 해양산업의 안전성 확보와 자율운항 선박의 상용화에 기여할 중요한 기술로, 지속적인 추후 연구와 실무 적용이 필요.

감사합니다