Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.33, NO.3, Jun. 2023 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2023.33.3.449

적대적 공격을 방어하기 위한 StarGAN 기반의 탐지 및 정화 연구*

박 성 준,[1†] 류 권 상,[2] 최 대 선[2‡]
[1,2]숭실대학교 (학생, 교수)

StarGAN-Based Detection and Purification Studies to Defend against Adversarial Attacks*

Sungjune Park,[1†] Gwonsang Ryu,[2] Daeseon Choi[2‡]
[1,2]Soongsil University (Undergraduate student, Professor)

요 약

인공지능은 빅데이터와 딥러닝 기술을 이용해 다양한 분야에서 삶의 편리함을 주고 있다. 하지만, 딥러닝 기술은 적대적 예제에 매우 취약하여 적대적 예제가 분류 모델의 오분류를 유도한다. 본 연구는 StarGAN을 활용해 다양한 적대적 공격을 탐지 및 정화하는 방법을 제안한다. 제안 방법은 Categorical Entropy loss를 추가한 StarGAN 모델에 다양한 공격 방법으로 생성된 적대적 예제를 학습시켜 판별자는 적대적 예제를 탐지하고, 생성자는 적대적 예제를 정화한다. CIFAR-10 데이터셋을 통해 실험한 결과 평균 탐지 성능은 약 68.77%, 평균 정화 성능은 약 72.20%를 보였으며 정화 및 탐지 성능으로 도출되는 평균 방어 성능은 약 93.11%를 보였다.

ABSTRACT

Artificial Intelligence is providing convenience in various fields using big data and deep learning technologies. However, deep learning technology is highly vulnerable to adversarial examples, which can cause misclassification of classification models. This study proposes a method to detect and purification various adversarial attacks using StarGAN. The proposed method trains a StarGAN model with added Categorical Entropy loss using adversarial examples generated by various attack methods to enable the Discriminator to detect adversarial examples and the Generator to purification them. Experimental results using the CIFAR-10 dataset showed an average detection performance of approximately 68.77%, an average purification performance of approximately 72.20%, and an average defense performance of approximately 93.11% derived from restoration and detection performance.

Keywords: Adversarial example, Generative Adversarial Networks, Adversarial defense, Purification network

I. 서 론

인공지능은 빅데이터와 딥러닝을 활용하여, 이미지 분류, 객체 탐지, 자연어 처리 등 많은 분야에서 활용되어 삶의 편리함을 준다. 하지만, 인공지능의 핵심 기술인 딥러닝은 많은 보안 취약점을 가지고 있

Received(03. 06. 2023), Modified(04. 24. 2023), Accepted(04. 24. 2023)
* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발)과 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1A2C1014813)
* 본 논문은 2022년도 한국정보보호학회 동계학술대회에 발표한 우수논문을 개선 및 확장한 것임.
† 주저자, joey25@soongsil.ac.kr
‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

어 딥러닝 보안 문제에 대한 관심이 증가하고 있다[1]. 이러한 딥러닝 모델은 공격자가 의도적으로 입력하는 적대적 예제에 대해 매우 취약하며, 일관적으로 오분류 하게 된다. 이미지에서의 적대적 예제는 사람의 눈으로는 식별하기 어려운 만큼의 노이즈가 추가된 이미지로 딥러닝 모델이 오분류를 하게 된다.

적대적 공격은 사람의 눈으로 식별하기 어려운 작은 노이즈가 추가된 이미지인 적대적 예제를 생성하여 딥러닝 모델의 입력 데이터에 넣어 모델의 오분류를 유도하는 공격이다. 적대적 공격은 적대적 예제가 특정 클래스로 오분류하도록 유도하는 표적 공격(Targeted Attack)과 단순히 원본 클래스 이외의 클래스로 오분류하도록 유도하는 무표적 공격(Untargeted Attack)으로 나뉜다[5]. 예를 들어, 동물을 인식하여 구분하는 DNN모델이 판다를 긴팔원숭이로 오인식하도록 적대적 예제를 생성하는 공격은 표적 공격이고, 판다를 판다 이외의 다른 동물로 오인식하도록 적대적 예제를 생성하는 공격은 무표적 공격이다.

지금까지의 적대적 예제에 대한 방어 방법으로는 하나의 공격 방법으로 적대적 예제를 생성하여 DNN(Deep Neural Networks)[2]모델 학습에 활용하는 적대적 학습(Adversarial Training)[7,17,18,19], 입력 이미지와 변환한 이미지에 대한 DNN 모델의 출력 값의 차이를 이용하는 탐지(Detection)[12,20,21,22], 한 가지 방법으로 적대적 예제를 생성하여 Purifier Network를 학습시키는 Filtering, Denoising, Purifier 등[14]이 있다.

본 논문에서는 다양한 공격 방법을 탐지 및 정화(Purification)를 위해 StarGAN[15]을 활용하고 기존의 정화 모델과 비교를 하려고 한다. 기존에 GAN(Generative Adversarial Networks)[3]을 사용하여 공격을 방어하기 위해서는 공격 종류당 하나의 GAN을 사용해야 하는 문제가 있었으나, 본 논문에서 제안하는 방법은 공격 종류가 여러 가지 일지라도 하나의 GAN만 사용하여 공격을 탐지 및 정화가 가능하도록 한다. 본 논문에서는 StarGAN의 판별자(Discriminator)는 탐지기로 사용하여 입력되는 적대적 예제를 탐지하고, 생성자(Generator)는 Purifier로 사용하여 탐지하지 못한 적대적 예제를 정화하여 딥러닝 모델이 올바른 클래스로 분류되도록 한다. 또한 기존의 정화 모델은 Denoising U-net[16] 구조를 가진 HGD(High-Level Representation Guided Denoiser)[4]를 사용한다.

본 논문에서 기여하는 바는 다음과 같다.

1. 다양한 적대적 공격에 대응하기 위해 하나의 모델로 탐지와 정화를 동시에 활용하는 방법을 제안한다.
2. 기존의 정화 모델인 HGD보다 본 논문에서 제시하는 방어방법의 방어성능이 약 14.53% 더 높다.
3. 모델의 파라미터 변화를 통해 적대적 예제에 더욱 강건한 모델을 탐색하였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 다루고, 3장에서는 제안 방법에 대한 설명을 한다. 4장에서는 제안 방법을 평가하기 위한 실험 및 결과에 대해 다루고 5장에서는 제안 방법에 대해 고찰하며, 마지막으로 6장에서 결론을 맺는다.

II. 관련 연구

2.1 적대적 공격 생성 기술

적대적 공격 생성 방법에는 FGSM, BIM, PGD, CWL2, DeepFool, MIM, 등[6-11]이 존재한다. FGSM(Fast Gradient Sign Method)[6] 공격 방법은 Goodfellow가 제안하였으며, 한 번의 변조를 통해 간단하고 빠르게 적대적 예제를 생성한다. FGSM공격 방법은 다음과 같이 적대적 예제를 생성한다.

$$\hat{x} = x - \epsilon \bullet sign(\nabla_x loss_{F,t}(x)) \qquad (1)$$

이 수식에서 ϵ은 최대 변조량 sign은 방향을 결정하여, DNN모델의 손실함수의 기울기를 이용해 적대적 예제를 생성한다.

이 방법은 공격의 생성 속도는 빠르지만, 공격 성공률이 높지 않다는 한계가 존재한다. Kurakin은 정해진 횟수만큼 FGSM을 반복하여 적대적 예제를 생성하는 I-FGSM(Iterative-FGSM)이라고도 불리는 BIM(Basic Iterative Method)[7] 공격 방법을 제안하였다. FGSM을 여러번 반복하였기 때문에 FGSM보다 모델에 대한 공격 성공률이 더 높다. MIM공격 방법은 다음과 같이 적대적 예제를 생성한다.

$$\hat{x}_i = \hat{x}_{i-1} - clip_{\epsilon}(\alpha \bullet sign(\nabla_{\hat{x}_{i-1}} loss_{F,t}(\hat{x}_{i-1}))) \quad (2)$$

이 수식에서 α는 반복마다의 변조량을 의미하고 clip은 변조량의 폭을 ϵ만큼으로 제한하는 함수이다.

Mardy가 제안한 PGD(Projected Gradient Descent)[8] 공격 방법은 BIM 공격과 동일하나 변형이 없는 원본 이미지에서 공격을 시작하는 BIM이나 FGSM과는 다르게 원본 이미지에 랜덤 노이즈를 추가해 공격을 시작하는 방법이다. CWL2(Carlini and Wagner)[9] 공격 방법은 Carlini와 Wagner가 제안한 Distance Metric인 L0, L2, L∞를 기반으로 한 3가지 적대적 공격 방법 중 가장 많이 사용되는 L2 공격 방법이다. CWL2공격 방법은 다음과 같이 적대적 예제를 생성한다.

$$min \| \| \hat{x} - x \| \|_2 + c \bullet f(\hat{x}) \quad (3)$$

여기서 $f(\hat{x})$는 다음과 같이 정의된다.

$$f(\hat{x}) = \max(\max\{Z(\hat{x})_i : i \neq t\} - Z(\hat{x}_t), -k) \quad (4)$$

여기서 k는 생성된 적대적 예제가 DNN모델이 목표 클래스인 t로 분류할 확률을 의미하고, $Z(\hat{x})$는 적대적 예제 $\hat{x}$에 대한 DNN모델의 출력 벡터를 의미한다.

DeepFool[10] 공격 방법은 Moosavi-Dezfooli가 제안한 방법으로 Euclidean distance를 최적화 하는 무표적 공격 방법이다. DeepFool은 최적화를 통해 FGSM에 비해 더 작은 노이즈를 가지지만, 복잡한 비선형 구조를 가지기 때문에 적대적 공격을 생성하는데 더 많은 시간이 걸린다.

Dong이 제안한 MIM[11]은 손실 함수를 이용해, 그 기울기 방향으로 가속도를 누적시켜 적대적 예제를 생성하는 방법으로, 블랙박스 공격에서 높은 공격 성공률을 달성하였다. MIM공격 방법은 다음과 같이 적대적 예제를 생성한다.

$$\hat{x}_{i+1} = \hat{x}_i + \alpha \bullet sign(g_{i+1}) \quad (5)$$

이 수식에서 α는 반복마다의 변조량을 의미하고 g_{i+1}는 기울기 방향으로 가속도를 누적시키는 함수이며 다음과 같다.

$$g_{i+1} = \mu \bullet g_i + \frac{\nabla_x J(\hat{x_i}, y)}{\| \nabla_x J(\hat{x_i}, y) \|_1} \quad (6)$$

이 수식에서 μ는 이전 단계에서 얻은 g_i의 반영 비율이고, y는 원본 이미지의 클래스를 의미한다.

2.2 적대적 공격 방어 기술

적대적 공격에 대한 방어 방법으로는 적대적 학습, 탐지, 정화가 있다. 적대적 학습은 Goodfellow[5]가 처음으로 제안하였으며, 적대적 공격에 대한 모델의 견고함(robustness)을 효과적으로 향상시키기 위한 방어 방법이다. 적대적 학습은 모델 손실 최대화 및 최소화의 원칙을 기반으로 하는데, 최대화 단계에서는 생성된 적대적 예제를 훈련 데이터셋에 작은 부분으로 넣어주고 최소화 단계에서는 손실 최소화를 위해 넣어준 적대적 예제를 이용하여 모델의 파라미터를 업데이트 시킨다. 이러한 적대적 학습은 다양한 적대적 공격에 대해 강력하도록 DNN모델을 훈련하지만, 기존에 훈련된 DNN모델보다 원본 이미지에 대한 분류 정확도가 낮다.

탐지는 적대적 예제의 특성과 정상 데이터의 특성 차이를 기반으로 적대적 예제를 탐지하는 방어 방법이다. 오토인코더를 이용해 입력 이미지와 정화된 입력 이미지의 차이를 통해 적대적 예제를 탐지하거나[12], 입력 이미지에 대한 robust feature를 추출해 입력이 미지에 대한 분류결과에서 예측되는 feature와 비교하여 탐지하는 방법도 있다[13].

정화는 적대적 예제에 존재하는 노이즈를 제거하는 방법이다. Samangouei[14]는 GAN을 이용해 적대적 공격의 노이즈를 줄이는 Defense-GAN을 제안하였는데, 이는 생성자를 생성 이미지와 원본 이미지의 차이가 최소가 되도록 학습하고, 생성자를 정상 이미지와 생성 이미지를 잘 구분하도록 학습한다. 정화 후에도 남은 적대적 예제의 노이즈가 모델의 오분류를 유도한다는 문제점을 해결하기 위해 Liao 등[4]은 하여 이미지를 정화 하는 방법인 HGD를 제안하였다. HGD는 오토인코더에 U-Net

구조를 활용한 오토인코더로, 정화된 이미지와 입력 이미지의 거리 차이를 줄이면서도 DNN모델이 정화된 이미지를 오분류 하지 않도록 원본 이미지와 노이즈가 제거된 이미지 사이의 차이로 정의된 손실 함수를 사용하였다. 이 외에도 적대적 공격을 방어하기 위한 다양한 방어 방법이 있다.

III. 제안 방법

3.1 StarGAN 모델

본 논문에서는 하나의 GAN으로 다양한 적대적 공격을 탐지하고 정화하기 위해 StarGAN 모델을 사용하였다. StarGAN의 생성자와 생성자는 다양한 공격 방법으로 만들어진 적대적 예제들을 Fig. 1.과 같은 방법으로 학습한다.

생성자의 학습 과정은 다음과 같다. 입력 이미지와 타겟 도메인 즉, 원본 이미지로의 도메인을 함께 입력받아 입력 이미지를 타겟 도메인으로 생성한다. 생성한 이미지를 다시 원래 도메인으로 복구하며 학습한다. 생성자의 학습 과정은 다음과 같다. 생성자에서 타겟 도메인으로 생성한 이미지와 원본 이미지를 입력받아 진짜 이미지인지 가짜 이미지인지 구분하며, 정상 이미지와 적대적 공격의 도메인을 분류하며 학습한다.

Inference 단계에서 생성자는 정상과 적대적 공격을 탐지하고, 동시에 이미지의 공격 방식을 분류한다. 생성자는 입력 이미지를 타겟 도메인으로 변환하는 역할을 하며 공격 이미지를 원본으로 정화한다. 여기서 도메인은 같은 attribute value를 공유하는 이미지들의 집합을 칭한다.

아래의 수식들에서 G는 생성자, D는 생성자로, D_{src}는 입력 이미지 x에 대해 생성자가 진짜 이미지인지 가짜 이미지인지를 판별한다. D_{cls}는 입력 이미지 x에 대해 생성자가 예측한 도메인과 실제 도메인을 비교한다. c는 바꾸고자 하는 도메인이고, c'는 입력 이미지의 원래 도메인이다.

StarGAN 모델에서 사용하는 생성자의 loss는 진짜 이미지인지 가짜 이미지인지 구별할 수 있도록 학습하게 하는 Adversarial loss로 다음과 같이 정의된다.

$$L_{adv} = E_x[\log D_{src}(x)] + E_{x,c}[\log(1 - D_{src}(G(x,c)))] \quad (7)$$

이미지가 원래 도메인으로 잘 분류될 수 있도록 하는 Domain Classification loss는 다음과 같이 정의된다.

$$L_{cls}^{r} = E_{x,c'}[-\log D_{cls}(c' \mid x)] \quad (8)$$

생성자의 Adversarial loss는 수식 (7)과 같고, 생성자에서 생성된 이미지를 타겟 도메인으로 잘 분류되게 하는 Domain Classification loss는 다음과 같이 정의된다.

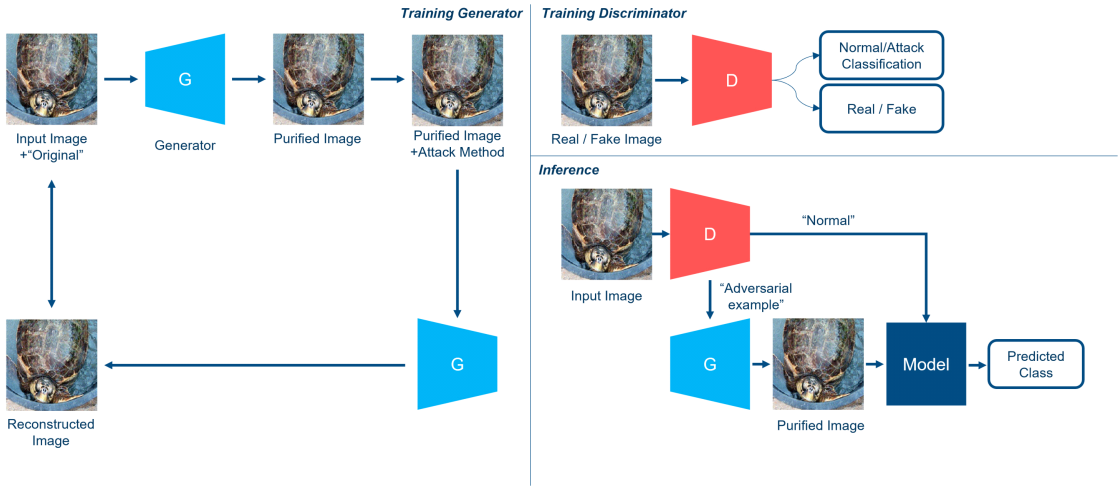

Fig. 1. Architecture of StarGAN-based Detection and Purification Model

$$L_{cls}^{f} = E_{x,c}[-\log D_{cls}(c \mid G(x,c))] \quad (9)$$

Reconstruction loss는 입력 이미지와 복구된 이미지의 맨하튼 거리 값을 갖는 loss로 다음과 같이 정의된다.

$$L_{rec} = E_{x,c,c'}[\|x - G(G(x,c),c')\|_1] \quad (10)$$

이에 더해 Categorical Cross Entropy loss를 추가해서 사용할 것이다. Categorical Cross Entropy loss를 추가한 StarGAN의 전체 loss 수식은 다음과 같다.

$$\begin{aligned} L_D &= -L_{adv} + \lambda_{cls} L_{cls}^{r} \\ L_G &= L_{adv} + \lambda_{cls} L_{cls}^{f} + \lambda_{rec} L_{rec} \\ &+ \lambda_{cce} L_{cce} \end{aligned} \quad (11)$$

3.2 Categorical Entropy

이미지를 타겟 도메인으로 변환하는 성능을 높이기 위해 다중 클래스 구분(Multi-class classification)에 도움이 되는 Categorical Cross Entropy loss를 StarGAN 학습 과정에 추가한다. 생성자에서 생성된 이미지를 타겟 모델인 VGG16에 넣어 예측 클래스와 실제 클래스의 값에 활성화 함수 Softmax를 적용하고 Cross-Entropy loss를 구한다. Categorical Cross Entropy loss의 수식은 다음과 같다.

$$L_{cce} = -\sum_{i=1}^{i=N} y_i \cdot \log(\hat{y_i}) \quad (12)$$

이 수식에서 N은 총 클래스의 개수이고 y_i는 원래 클래스, $\hat{y_i}$는 예측한 클래스를 의미한다.

IV. 실 험

4.1 실험 환경

본 논문에서 제안한 적대적 예제 탐지 및 정화 방법을 학습 및 평가하기 위해 CIFAR-10 데이터 셋을 사용하였다. 이미지 분류 모델은 VGG16을 CIFAR-10 데이터셋으로 학습하여 사용하였고 이 모델의 원본 이미지에 대한 분류 정확도는 약 90.60%이다. StarGAN 학습에 사용한 적대적 예제는 FGSM, CWL2, DeepFool, PGD 공격 방법을 사용해서 생성하였다. 평가에 사용한 적대적 예제는 BIM, CWL2, DeepFool, FGSM, MIM, PGD 공격 방법을 이용해 생성하였다.

BIM 공격에서의 최대변조량은 0.031, 반복마다의 변조량은 0.007, 반복횟수는 10으로 설정하였고 CWL2 공격에서의 이진탐색횟수는 9로 설정하였다. DeepFool 공격에서의 최대반복횟수는 100, 공격 후보 class갯수는 10으로 설정하였고 FGSM 공격에서의 최대변조량은 0.031로 설정하였다. MIM 공격에서의 최대변조량은 0.031, 반복마다의 변조량은 0.007, 반복횟수는 10으로 설정하였고 PGD 공격에서의 최대변조량은 0.031, 반복마다의 변조량은 0.007, 반복횟수는 10으로 설정하였다.

StarGAN 모델의 학습은 원본 이미지와 FGSM, CWL2, DeepFool, PGD를 학습한 모델과 원본 이미지, DeepFool, PGD를 학습한 모델을 실험에 사용하였다. HGD 모델은 원본 이미지, FGSM, CWL2, DeepFool, PGD를 학습한 모델을 실험에 사용하였다.

학습 데이터셋은 이미지 분류 모델이 잘 분류한 데이터들을 적대적 예제로 변환 후 클래스 분류 모델이 오분류한 데이터만을 학습에 사용했으며, 총 290,886장의 데이터를 사용하였다. 테스트 데이터 셋은 이미지 분류 모델이 잘 분류한 데이터들을 적대적 예제로 변환하여 사용했으며 총 63,420장의 데이터를 사용하였다.

4.2 평가 방법

적대적 예제를 가장 잘 방어하는 최적의 파라미터를 찾기 위해 파라미터를 조절하였다. 파라미터 조절을 통한 비교를 위해, Loss 수식에서 λ_{cls}는 1.0, λ_{rec}는 10.0 λ_{cce}는 2.0와 λ_{cls}는 1.0, λ_{rec}는 10.0 λ_{cce}는 8.0와 λ_{cls}는 1.0, λ_{rec}는 5.0 λ_{cce}는 2.0로 설정하였다. 각각의 파라미터는 Loss 수식에서 각각의 Loss에 대해 어느정도 가중치를 둘지 정하는 파라미터이다.

원본 이미지에 대한 탐지율은 원본 이미지를 원본 이미지로 분류한 개수 / 총 원본 이미지 개수로 구하

였고 적대적 예제는 적대적 예제를 적대적 예제로 분류한 개수 / 총 적대적 예제 개수로 구하였다. 또한 학습되지 않은 적대적 공격은 추론단계에서 생성자가 정상이 아니라고 분류했으면 적대적 공격 탐지에 성공하였다고 보았다. 정화률은 이미지 분류 모델인 VGG16에 생성자가 정화한 이미지를 넣었을 때 원래 라벨로 분류한 개수 / 총 이미지 개수로 구하였다. 방어 성능은 적대적 공격 탐지율에 탐지하지 못한 공격 중 정상 이미지로 정화되는 이미지의 비율을 더해서 구하였다.

파라미터 조절을 통한 적대적 공격 탐지, 정화, 방어 성능은 Table 1.과 Table 2.와 같다. Table 1.은 Original, FGSM, PGD, CWL2, DeepFool을 학습하였을 때 방어 성능이며, Table 2.는 Original, PGD, DeepFool을 학습하였을 때 방어 성능을 나타낸다. 표에서 DR(Detection Rate)은 탐지 성능, PR(Purification Rate)은 정화 성능, Total은 총 방어 성능을 의미한다. 생성한 적대적 예제 중 공격이 성공한 적대적 예제를 대상으로 성능을 평가하였다.

4.3 실험 결과

Table 1.의 탐지 성능에서 원본 이미지의 탐지성능이 가장 높은 모델은 파라미터가 λ_{rec}는 10.0 λ_{cce}는 2.0인 모델이 66.91%로 가장 잘 탐지하였고, 적대적 공격에 대한 탐지 성능은 파라미터가 λ_{rec}는 5.0 λ_{cce}는 2.0인 모델이 가장 잘 탐지하였다. 복원 성능으로는 파라미터가 λ_{rec}는 10.0 λ_{cce}는 2.0인 모델이 BIM, PGD에 대해 성능이 각각 70.22%, 70.08%로 가장 잘 복원하였고, 파라미터

Table 1. Model Evaluation of Proposed Method Trained with Original, FGSM, DeepFool, PGD, and CWL2

Input	StarGAN(Train data = Original, FGSM, PGD, CWL2, DeepFool)								
	$\lambda_{rec}=10$, $\lambda_{cce}=2$			$\lambda_{rec}=10$, $\lambda_{cce}=8$			$\lambda_{rec}=5$, $\lambda_{cce}=2$		
	DR	PR	Total	DR	PR	Total	DR	PR	Total
Original	66.91%	85.70%	95.27%	24.47%	70.31%	77.58%	16.81%	86.57%	88.83%
BIM	66.33%	70.22%	89.97%	86.35%	67.46%	95.56%	87.93%	69.37%	96.30%
CWL2	33.63%	80.74%	87.22%	76.22%	66.13%	91.95%	80.54%	81.38%	96.38%
DeepFool	36.69%	80.13%	87.42%	78.25%	65.34%	92.46%	81.70%	80.78%	96.48%
FGSM	82.69%	64.95%	93.93%	87.88%	66.86%	95.98%	88.76%	65.06%	96.07%
MIM	82.61%	60.94%	93.21%	87.88%	66.37%	95.92%	88.74%	59.64%	95.46%
PGD	66.39%	70.08%	89.94%	86.38%	67.22%	95.54%	87.96%	69.34%	96.31%

Table 2. Model Evaluation of Proposed Method Trained with Original, DeepFool, PGD

Input	StarGAN(Train data = Original, PGD, DeepFool)								
	$\lambda_{rec}=10$, $\lambda_{cce}=2$			$\lambda_{rec}=10$, $\lambda_{cce}=8$			$\lambda_{rec}=5$, $\lambda_{cce}=2$		
	DR	PR	Total	DR	PR	Total	DR	PR	Total
Original	63.16%	87.37%	95.35%	72.14%	63.01%	89.69%	50.80%	85.76%	92.99%
BIM	82.49%	65.61%	93.98%	64.41%	60.92%	86.09%	82.65%	66.93%	94.26%
CWL2	39.28%	83.08%	89.73%	30.85%	59.06%	71.69%	47.16%	80.66%	89.78%
DeepFool	42.20%	82.38%	89.82%	32.13%	58.37%	71.75%	48.90%	79.60%	89.58%
FGSM	85.95%	63.60%	94.89%	81.35%	60.81%	92.69%	86.68%	64.14%	95.22%
MIM	85.88%	57.73%	94.03%	81.03%	60.48%	92.50%	86.56%	58.84%	94.47%
PGD	82.41%	65.63%	93.95%	64.33%	60.85%	86.04%	82.74%	67.13%	94.33%

가 λ_{rec}는 10.0 λ_{cce}는 8.0인 모델은 FGSM과 MIM에 대해 성능이 각각 66.86%, 66.37%로 가장 성능이 높았다. 원본 이미지와 CWL2, DeepFool에 대한 복원 성능은 파라미터가 λ_{rec}는 5.0 λ_{cce}는 2.0인 모델이 가장 성능이 높았다. 총 방어 성능에서는 파라미터가 λ_{rec}는 10.0 λ_{cce}는 2.0인 모델이 원본 이미지에 대해 성능이 95.27%로 가장 높았고, 파라미터가 λ_{rec}는 5.0 λ_{cce}는 2.0인 모델인 모델이 모든 적대적 공격에 대해 성능이 가장 높았다.

Table 2.의 탐지성능에서 원본 이미지에 대한 탐지 성능이 가장 높은 모델은 파라미터가 λ_{rec}는 10.0 λ_{cce}는 8.0인 모델이 72.14%로 가장 높은 탐지 성능을 보였다. 적대적 공격에 대한 탐지 성능은 파라미터가 λ_{rec}는 5.0 λ_{cce}는 2.0인 모델이 가장 잘 탐지하였다. 복원 성능으로는 파라미터가 λ_{rec}는 10.0 λ_{cce}는 2.0인 모델이 원본이미지와 CWL2, DeepFool에 대해 성능이 각각 87.37%, 83.08%, 82.38%로 가장 잘 복원하였고, 파라미터가 λ_{rec}는 10.0 λ_{cce}는 8.0인 모델은 FGSM과 MIM에 대해 성능이 각각 60.81%, 60.48%로 가장 성능이 높았다. PGD에 대해서는 파라미터가 λ_{rec}는 5.0 λ_{cce}는 2.0인 모델이 가장 성능이 높았다. 총 방어 성능에서는 파라미터가 λ_{rec}는 10.0 λ_{cce}는 2.0인 모델이 원본 이미지와 DeepFool에 대해 각각 95.35%, 89.82%로 가장 높았고, 파라미터가 λ_{rec}는 5.0 λ_{cce}는 2.0인 모델이 BIM, CWL2, FGSM, MIM, PGD에 대해 가장 방어 성능이 높게 나왔다.

실험 결과에서 제일 성능이 좋게 나온 파라미터는 λ_{rec}는 10.0 λ_{cce}는 2.0이며 학습 데이터를 원본이미지와 PGD, DeepFool로 두었을 때 가장 성능이 좋았다. Table 1.에 λ_{rec}는 5.0 λ_{cce}는 2.0인 모델은 원본 이미지에 대한 탐지율이 현저히 낮기 때문에 방어 성능이 높게 나왔더라도 성능이 좋다고 할 수 없어 선정에서 제외하였다. 또한 오탐율은 학습 및 평가에 사용한 데이터가 분류 모델이 오탐하지 않

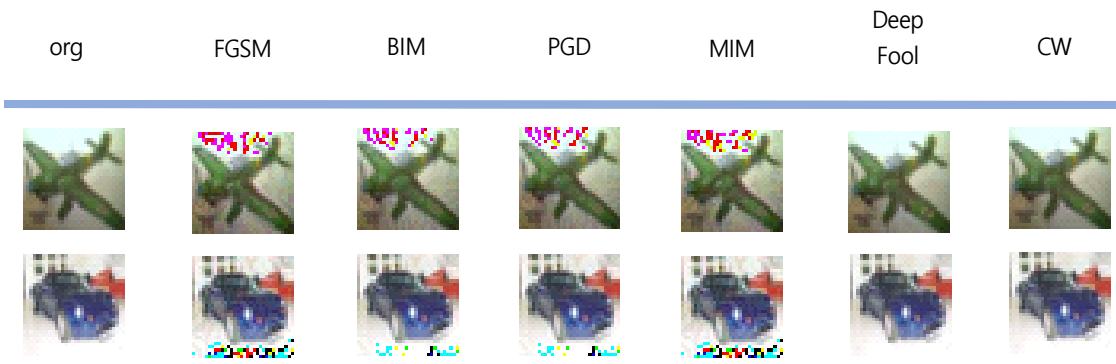

Fig. 2. Example of Adversarial Example

Table 3. Evaluation of Proposed Method for Adversarial Example

Input	StarGAN(Train data = Original, PGD, DeepFool) $\lambda_{rec}=10, \lambda_{cce}=2$			HGD
	DR	PR	Total	Total
Original	63.16%	87.37%	95.35%	87.30%
BIM	82.49%	65.61%	93.98%	78.43%
CWL2	39.28%	83.08%	89.73%	84.04%
DeepFool	42.20%	82.38%	89.82%	82.15%
FGSM	85.95%	63.60%	94.89%	69.96%
MIM	85.88%	57.73%	94.03%	69.60%
PGD	82.41%	65.63%	93.95%	78.54%

은 데이터를 사용하였으므로 Table 3.에서 100%-DR인 약 37%이며, 오탐하였을 경우 정화 모델에서 정화시키므로 총 95.35% 정도의 방어 성능을 보인다.

4.4 기존 연구와의 성능 비교

파라미터 조절을 통한 성능 비교를 하였을 때, 학습데이터가 Original, PGD, DeepFool이고 파라미터가 λ_{cls}는 1.0, λ_{rec}는 10.0 λ_{cce}는 2.0인 모델이 높은 성능을 보였다. Table 3.은 기존 연구인 HGD와 성능을 비교한 표이다. 모든 입력 데이터에 대해 방어 성능이 뛰어났고,특히 FGSM과 MIM에 대한 방어 성능이 약 24% 증가한 것을 보였다. 그 뒤를 이어 BIM과 PGD에 대한 방어 성능도 약 15%정도 뛰어났다. 원본 데이터에 대한 방어 성능은 약 8%정도 높고, DeepFool과 CWL2에 대한 방어 성능은 각각 7.67%, 5.69% 높다. 모든 입력 데이터에 대한 평균 방어 성능의 차이는 약 14.53%이다.

V. 고 찰

본 논문은 적대적 예제를 방어하기 위한 StarGAN 기반 탐지 및 정화 방법을 제안하였다. 공격 종류가 여러 가지 일지라도 하나의 GAN만 사용하여 공격을 탐지 및 정화가 가능하도록 하였다. 실험에서는 공격 종류를 4가지와 2가지로 각각 설정하여 StarGAN을 학습한 결과, 2가지일 때의 방어 성능이 더 높은 결과를 보였다. 하지만 원본 이미지에 대한 탐지율은 공격 종류를 4가지로 하였을 때 더 높은 탐지율을 보였고, BIM, FGSM, PGD, MIM 공격에 대한 정화율도 더 높은 결과를 보였다. 또한 정화율 자체는 HGD가 대체적으로 모든 StarGAN모델에 비해 더 높은 성능을 보이는 한계가 있다.

적대적 예제의 탐지 및 정화에서, BIM과 PGD는 FGSM을 여러번 반복한 적대적 공격이기에 StarGAN모델이 FGSM보다 탐지율이 낮은 결과를 보였고, 이 둘의 탐지율과 정화율은 유사한 결과를 보였다. DeepFool과 CWL2는 둘 다 L_2기반 적대적 공격이기에 비슷한 탐지율과 정화율을 보였고, 다른 공격 방법에 비해 낮은 탐지율을 보였지만 높은 정화율을 보였다.

모델의 파라미터를 조절하였을 때, 원본 데이터에 대한 탐지율이 낮아지는 경향을 보였다. 이는 생성자가 진짜와 가짜 이미지를 구분하는 역할과 원본과 공격 종류를 분류하는 2가지 역할을 하는데 생성자는 마지막 레이어를 2개로 나눠 각 역할에 맞게 학습한다. 두 개의 역할은 마지막 레이어 전까지의 생성자의 구조를 공유하므로 GAN을 학습시킬수록 생성자의 진짜와 가짜를 구분하는 역할의 성능은 떨어지고, 앞단의 레이어들을 공유하기 때문에 공격 종류를 구분하는 능력도 떨어진다.

VI. 결 론

본 논문에서는 하나의 GAN으로 다양한 도메인으로 이미지 변환 및 학습이 가능한 StarGAN 모델에 Categorical Cross Entropy loss를 추가하여 다양한 적대적 공격을 방어할 수 있는 방법을 제안하였다. 제안한 방법을 평가한 결과, 학습한 적대적 공격의 종류에 따른 방어율을 비교해 보았을 때, 4가지 적대적 공격을 학습한 StarGAN모델 보다 적대적 공격의 종류가 적은 2가지 적대적 공격을 학습한 StarGAN모델의 방어 성능이 더 높게 나타나는 결과를 보였다. 또한 기존의 정화 모델인 HGD와 비교하였을 때, 논문에서 제안한 방법이 적대적 예제에 대해 더 높은 방어 성능을 보였다.

향후 연구로는 더 강건한 모델을 구축하기 위해 FGSM과 BIM, MIM, PGD에 대한 정화 성능을 높이고, L_2기반 적대적 공격인 DeepFool과 CWL2에 대한 탐지 성능을 높이는 연구를 진행할 예정이다.

References

[1] G. Ryu and D. Choi, "A Research-Trends in Artificial Intelligence SecurityAttacks and Countermeasures," Review of KIISC, 30(5), pp. 93-99, Oct. 2020.

[2] SZE, Vivienne, et al. Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 2017.

[3] Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 (2020): pp. 139-144.

[4] Liao, Fangzhou, et al. "Defense against adversarial attacks using high-level representation guided denoiser." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[5] Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." Ieee Access 6 (2018): pp. 14410-14430.
[6] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
[7] Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." Artificial intelligence safety and security. Chapman and Hall/CRC, 2018. pp. 99-112.
[8] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
[9] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." 2017 ieee symposium on security and privacy (sp). Ieee, 2017.
[10] Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[11] Dong, Yinpeng, et al. "Boosting adversarial attacks with momentum." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[12] Meng, Dongyu, and Hao Chen. "Magnet: a two-pronged defense against adversarial examples." Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. 2017.
[13] Freitas, Scott, et al. "Unmask: Adversarial detection and defense through robust feature alignment." 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020.
[14] Samangouei, Pouya, Maya Kabkab, and Rama Chellappa. "Defense-gan: Protecting classifiers against adversarial attacks using generative models." arXiv preprint arXiv:1805.06605 (2018).
[15] Choi, Yunjey, et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[16] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.
[17] Ryu, Gwonsang, and Daeseon Choi. "Feature-based adversarial training for deep learning models resistant to transferable adversarial examples." IEICE TRANSACTIONS on Information and Systems 105.5 (2022): pp. 1039-1049.
[18] Ryu, Gwonsang, and Daeseon Choi. "A hybrid adversarial training for deep learning model and denoising network resistant to adversarial examples." Applied Intelligence (2022): pp. 1-14.

[19] Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." International conference on machine learning. PMLR, 2019.
[20] Xu, Weilin, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks." arXiv preprint arXiv:1704.01155 (2017).
[21] Ma, Shiqing, et al. "Nic: Detecting adversarial samples with neural network invariant checking." 26th Annual Network And Distributed System Security Symposium (NDSS 2019). Internet Soc, 2019.
[22] Ye, Dengpan, et al. "Detection defense against adversarial attacks with saliency map." International Journal of Intelligent Systems 37.12 (2022): pp. 10193-10210.

〈저자소개〉

박 성 준 (Sungjune Park) 학생회원
2019년 3월~현재: 숭실대학교 소프트웨어학부 학사과정
〈관심분야〉 인공지능 보안, 생성 AI

류 권 상 (Gwonsang Ryu) 정회원
2016년 2월: 공주대학교 응용수학과 학사
2018년 2월: 공주대학교 융합과학과 석사
2018년 3월~2020년 8월: 공주대학교 융합과학과 박사과정
2020년 9월~2022년 2월: 숭실대학교 융합소프트웨어학과 박사
2022년 3월~현재: 숭실대학교 사이버보안연구센터 연구교수
〈관심분야〉 인증, 이상거래탐지, 인공지능 보안

최 대 선 (Daeseon Choi) 종신회원
1995년 2월: 동국대학교 컴퓨터공학과 학사
1997년 2월: 포항공과대학교 컴퓨터공학과 석사
2009년 1월: 한국과학기술원 전산학과 박사
1997년 1월~1999년 6월: 현대정보기술 선임
1999년 7월~2015년 8월: 한국전자통신연구원 인증기술연구실 실장/책임연구원
2015년 9월~2020년 8월: 공주대학교 의료정보학과 부교수
2020년 9월~현재: 숭실대학교 소프트웨어학부 교수
2016년~현재: 정보보호학회 이사
〈관심분야〉 AI 보안, 인증, 개인정보보호, 이상거래탐지, 의료정보보안, 머신러닝