JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 2, pp. 31-41, February 2024
https://doi.org/10.9708/jksci.2024.29.02.031

A Study on the Efficacy of Edge-Based Adversarial Example Detection Model: Across Various Adversarial Algorithms

Jaesung Shim*, Kyuri Jo*

*Student, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea
*Professor, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea

[Abstract]

Deep learning models show excellent performance in tasks such as image classification and object detection in the field of computer vision, and are used in various ways in actual industrial sites. Recently, research on improving robustness has been actively conducted, along with pointing out that this deep learning model is vulnerable to hostile examples. A hostile example is an image in which small noise is added to induce misclassification, and can pose a significant threat when applying a deep learning model to a real environment. In this paper, we tried to confirm the robustness of the edge-learning classification model and the performance of the adversarial example detection model using it for adversarial examples of various algorithms. As a result of robustness experiments, the basic classification model showed about 17% accuracy for the FGSM algorithm, while the edge-learning models maintained accuracy in the 60-70% range, and the basic classification model showed accuracy in the 0-1% range for the PGD/DeepFool/CW algorithm, while the edge-learning models maintained accuracy in 80-90%. As a result of the adversarial example detection experiment, a high detection rate of 91-95% was confirmed for all algorithms of FGSM/PGD/DeepFool/CW. By presenting the possibility of defending against various hostile algorithms through this study, it is expected to improve the safety and reliability of deep learning models in various industries using computer vision.

▸ **Key words**: Deep Learning Model, Computer Vision, Convolutional Neural Network, Adversarial Example, Edge-based Classification, Adversarial defense

• First Author: Jaesung Shim, Corresponding Author: Kyuri Jo
*Jaesung Shim (climbsky@chungbuk.ac.kr), Dept. of Computer Engineering, Chungbuk National University
*Kyuri Jo (kyurijo@chungbuk.ac.kr), Dept. of Computer Engineering, Chungbuk National University
• Received: 2024. 01. 11, Revised: 2024. 02. 07, Accepted: 2024. 02. 07.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

딥러닝 모델(Deep Learning Model)은 컴퓨터 비전(Computer Vision) 분야의 이미지(Image) 분류 및 객체 탐지와 같은 작업에서 뛰어난 성과를 보이며, 실제 산업 현장에서 다양하게 활용되고 있다. 최근 다양한 알고리즘(Algorithm)의 적대적 예제를 이용하여 딥러닝 모델의 취약성을 지적하며, 강건성 향상 방안을 제시하는 연구들이 활발하게 진행되고 있다. 적대적 예제는 오분류를 유도하기 위해 작은 노이즈(Noise)가 추가된 이미지로서, 딥러닝 모델을 실제 환경에 적용 시 중대한 위협이 될 수 있다. 본 논문에서는 다양한 알고리즘의 적대적 예제를 대상으로 에지 학습 분류 모델의 강건성 및 이를 이용한 적대적 예제 탐지 모델의 성능을 확인하고자 하였다. 강건성 실험 결과, FGSM(Fast Gradient Sign Method) 알고리즘에 대하여 기본 분류 모델이 약 17%의 정확도를 보였으나, 에지(Edge) 학습 모델들은 60~70%대의 정확도를 유지하였고, PGD(projected gradient descent)/DeepFool/CW(Carlini-Wagner) 알고리즘에 대해서는 기본 분류 모델이 0~1%의 정확도를 보였으나, 에지 학습 모델들은 80~90%의 정확도를 유지하였다. 적대적 예제 탐지 실험 결과, FGSM/PGD/DeepFool/CW의 모든 알고리즘에 대해서 91~95%의 높은 탐지율을 확인할 수 있었다. 본 연구를 통하여 다양한 적대적 알고리즘에 대한 방어 가능성을 제시함으로써, 컴퓨터 비전을 활용하는 여러 산업 분야에서 딥러닝 모델의 안전성 및 신뢰성 제고를 기대한다.

▸ **주제어**: 딥러닝 모델, 컴퓨터 비전, 합성곱 신경망, 적대적 예제, 에지 기반 분류, 적대적 방어

I. Introduction

컴퓨터 비전 분야에서 뛰어난 성능을 나타내고 있는 딥러닝 모델은 이미지 분류 및 객체 탐지 기능을 활용하는 다양한 산업 분야에 적용되어 큰 성과를 나타내고 있다. 그런데 최근 이러한 딥러닝 모델이 적대적 예제에 취약하다는 문제가 지속해서 제기되고 있다[1-6]. 적대적 예제란 사람의 눈에는 보이지 않는 미세한 변형(perturbation)을 통해 딥러닝 모델의 예측이 잘못되도록 유도하는 이미지를 말한다[1]. 적대적 예제는 현실 세계에서의 딥러닝 모델의 실제 활용에 심각한 위협이 될 수 있는데, 예를 들어 적대적 예제를 이용한 공격을 통하여 자율주행 자동차의 충돌 사고나 보안 시스템의 우회 침입 등을 유발할 수 있다[7-9].

적대적 예제 공격을 방어하기 위한 다양한 연구가 진행되고 있는데[12], 대표적인 방법으로 딥러닝 모델을 적대적 예제에 노출하여 학습시키는 적대적 훈련이 있다[1][7][10-11]. 그러나 적대적 훈련은 모델의 성능(분류 정확도) 저하, 고비용의 계산 과정, 새로운 방식의 적대적 예제에는 적용이 불가하다는 단점들이 있다[13].

적대적 예제에 대한 분류 모델의 강건성을 향상하기 위하여 Borji는 원본 이미지에 객체의 에지를 더하여 학습시키는 방법을 제안하였으며[14], Xu et al.은 적대적 예제를 탐지하는 방법으로서 Feature Squeezing 모델과의 예측값 비교 방법을 제안하였다[15].

본 논문의 선행 연구[16]에서는, Borji와 Xu et al.의 논문을 기반으로, 에지 모델과 일반 모델 간의 예측값 비교를 통하여 FGSM 알고리즘을 적용하여 생성된 적대적 예제를 높은 정확도로 탐지할 수 있었다. 표 1은 각 epsilon(eps) 별 FGSM 적대적 예제(Adv. Image)와 클린 이미지(Cln. Image)를 섞은 테스트 데이터에 대하여 적대적 예제와 클린 이미지를 예측한 결과이다.

Table 1. Detection Results for Adversarial Examples[16]

eps	Adv. Image		Cln. Image	
	Correct	Wrong	Correct	Wrong
0.02	85.47%	14.53%	91.65%	8.35%
0.05	84.64%	15.36%	90.74%	9.26%
0.1	91.44%	8.56%	88.10%	11.90%
0.2	95.47%	4.53%	88.02%	11.98%
0.3	87.61%	12.39%	88.27%	11.73%

선행 연구에서는 FGSM 알고리즘만을 적용하여 탐지 모델의 성능을 확인하였기 때문에, 다른 알고리즘의 적대적 예제에 대해서도 동일한 탐지 성능이 보장될 수 있는지가 의문이었다. 이에, 본 후속 연구를 통하여, 다양한 알고리즘의 적대적 예제들을 대상으로 에지를 이용한 탐지 모델의 성능을 확인하고, 그 탐지 결과를 비교하고자 한다.

본 연구의 기여점은 다음과 같다.

1. 에지 학습으로 적대적 예제에 대한 분류 정확도가 향상됨을 입증함으로써, 적대적 예제에 대한 분류 모델의 강건성 강화 방안을 제시하였다.

2. 다양한 적대적 알고리즘에 대한 탐지 모델의 일관성 있는 우수한 탐지 결과를 통하여, 적대적 예제에 대한 효과적인 방어 메커니즘을 제시하였다.

3. 결과적으로 본 연구를 통하여, 다양한 산업에서 폭넓게 활용되고 있는 컴퓨터 비전 분야 딥러닝 모델의 신뢰성이 향상될 수 있을 것이다.

본 논문의 2장에서는 주요 적대적 예제 생성 알고리즘 및 적대적 예제 방어 기법과 이미지 내 에지 탐지 방식들을 소개하고, 3장에서는 본 연구의 적대적 예제 탐지 모델을 설명한 후, 4장에서 상세한 실험 과정 및 결과를 설명하고, 마지막으로 5장에서 결론을 제시하였다.

II. Preliminaries

1. Related works

1.1 Adversarial Example

컴퓨터는 이미지를 인식할 때, 이미지를 구성하는 각 픽셀(Pixel)에 대하여 빛의 삼원색인 빨강(Red), 초록(Green), 파랑(Blue)에 대한 값으로 인식하며, 각각의 RGB 값은 8 bits(2^8, 즉 256개)의 정보로 표현된다. 해당 정보를 인식할 때 사람은 미세한 변화를 인지할 수 없지만, 딥러닝 모델에서는 미세한 변화가 여러 계층으로 이루어진 복잡한 네트워크를 통과하면서 가중치 및 차원의 증가로 인하여 증폭되어 결과적으로 이미지를 다르게 인식하게 된다[1]. 적대적 예제는 이러한 딥러닝 모델의 특성을 이용하여, 클린 이미지(Clean Image)에 의도적으로 작은 노이즈를 삽입함으로써, 딥러닝 모델이 오분류를 일으키도록 생성된 이미지이다.

주요 적대적 예제 생성 기법들은 다음과 같다.

- FGSM(Fast Gradient Sign Method) : Ian Goodfellow 등이 2014년에 발표한 적대적 예제 생성 기법으로, 분류 모델의 손실 함수를 미분하여 계산된 기울기의 부호를 이용하여 미세한 노이즈를 이미지에 삽입한다[1]. 사람은 인지하기 어려운 작은 노이즈를 클린 이미지에 삽입하여 모델의 오분류를 유도하는, 매우 간단하고도 효과적인 기법이다.

- PGD(Projected Gradient Descent) : Aleksander Madry 등이 2017년에 제안하였으며, FGSM이 적대적 예제를 단일 스텝(Step)으로 생성하는 것과는 달리, PGD는 스텝을 반복함으로써 더 강력한 적대적 예제를 생성하는 기법이다[2]. 스텝의 반복 횟수 및 학습률에 따라 기울기를 조금씩 업데이트함으로써 공격 성공률이 높은 적대적 예제를 생성할 수 있으나, 이를 위해 많은 자원이 필요하다.

- DeepFool : Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard가 2016년에 제안하였으며, L2 Norm으로 계산하였을 때 분류 모델이 예측한 클래스와 가장 가까운 초평면(클래스 간 경계)으로 클린 이미지를 이동시키는 작업을 반복적으로 수행하여 클래스의 오분류를 유도하는 방식이다[3]. 최소 크기의 노이즈 벡터(Vector)를 찾아 최적의 적대적 예제를 생성할 수 있다는 장점이 있으나, 이를 위해 많은 자원이 필요하다.

- C&W(Carlini & Wagner) : Nicholas Carlini와 David Wagner가 2017년에 제안하였으며, 다양한 설정을 통하여 강력한 적대적 예제를 생성할 수 있다[4]. 손실 함수 최적화를 통하여, 클린 이미지와 적대적 예제의 차이를 최소화하는 동시에 목표(오분류) 클래스에 대한 예측 확률을 최대화한다. FGSM이나 PGD 알고리즘과 달리 L2, L0, L∞ 공격이 모두 가능하며 성공률 또한 매우 우수하여, DNN(Deep Neural Network) 분류 모델의 강건성 평가를 위하여 많이 사용되고 있다.

- ZOO(Zeroth Order Optimization) : Pin-Yu Chen 등이 2017년에 제안하였으며, 분류 모델의 모든 정보를 알고 있는 상태에서 공격하는 화이트 박스 공격과 달리, 입력값과 출력값만으로 공격이 가능한 기법이다[5]. 입력값의 변화에 따른 분류 모델 출력값의 변화를 관찰하여 기울기를 추정하며, 이를 통해 원본 이미지에 추가할 최소 노이즈를 계산한다. 이렇게 생성된 적대적 예제는 전이성(Transferability)을 가지고 있어서 구조가 비슷한 다른 분류 모델에 대한 공격이 가능하다.

1.2 Adversarial Defense

적대적 예제에 대한 방어 기법들 가운데 본 연구와 관련된 주요 방어 기법으로 적대적 훈련 방법과 적대적 예제 탐지 방법이 있는데, 적대적 훈련은 직접적으로 분류 모델의 강건성을 향상하는 방법이고, 적대적 예제 탐지는 이미지 분류 전 적대적 예제 여부를 탐지하는 방법이다.

- 적대적 훈련(Adversarial Training) : 딥러닝 모델이 적대적 예제들을 직접 학습하여 사람이 인식하는 클래스로 예측하도록 하는 방법이다[1-2][10-11]. 적대적 예제 생성 네트워크(GAN, Generative Adversarial Networks)를

통하여 적대적 예제를 생성한 후 딥러닝 모델이 학습하게 함으로써 적대적 예제를 올바른 클래스로 분류하도록 하는 원리이다.

적대적 훈련은 알려진(학습한) 적대적 예제에 대하여 효과적인 방어 수단이지만, 적대적 예제의 생성 및 학습을 위해 많은 자원이 필요하며, 클린 이미지에 대한 분류 정확도가 낮아지는 문제, 그리고 학습하지 못한 알고리즘의 적대적 예제에 대한 방어가 어렵다는 단점이 있다.

- 입력 변환(Input Transformation) : 이미지를 모델에 입력하기 전에 변환 작업을 통하여 공격을 어렵게 만드는 방법이다[17-21]. 데이터의 차원을 축소하여 공격 표면을 감소시키거나[17-18], 적대적 예제의 노이즈를 직접 제거하는 방법[19-21] 등이 있다.

입력 변환 기법은 모델 구조를 변경하지 않아도 되며, 다양한 공격 유형에 대한 방어가 가능하고, 모델에 쉽게 적용할 수 있다는 장점이 있다. 그러나, 변환 과정에서 중요 정보가 손실되는 경우 모델의 성능이 저하될 수 있고, 입력 변환 기법을 우회 공격할 수 있으며[6], 변환 과정이 복잡할 경우 계산 비용이 증가하는 단점 또한 존재한다.

- 적대적 예제 탐지(Adversarial Detection) : 이미지에 대한 클래스 분류를 하기 전에, 클린 이미지인지 적대적 예제인지를 먼저 판단하도록 하는 방법이다[15][22-23]. 판단 결과, 클린 이미지라면 분류 예측 결과를 출력하지만, 적대적 예제라면 입력 이미지가 적대적 예제라는 탐지 결과를 출력한다.

탐지 기법 중 예측 불일치(prediction inconsistency) 기법은 어떤 적대적 예제가 모든 딥러닝 분류 모델을 속이기는 어렵다는 가정을 전제로, 여러 모델에 이미지를 입력하여 예측 결과의 일치 여부를 확인하는 기법[15][22]으로서, 성능 면에서 효과가 좋으며 자원이 적게 드는 장점이 있다. 객체의 에지를 이용한 탐지 모델[16]도 여기에 속한다.

1.3 Edge Detection

이미지를 인식할 때 사람은 이미지 내 객체의 형태(에지)에 더 의존하지만, 딥러닝 모델은 객체의 텍스처(Texture)에 더 편향된 인식을 한다[14]. 이것은 딥러닝 모델이 정보를 획득하는 픽셀들이 객체의 에지보다 텍스처에 더 많이 분포되어 있기 때문이며, 결과적으로 적대적 공격의 기회를 제공하게 된다. 만약, 딥러닝 모델이 이미지 내 객체의 에지 부분을 강조하여 학습한다면 적대적 예제에 더 강건해질 수 있을 것이다[14].

에지 탐지(Edge Detection)는 이미지의 픽셀값이 급격하게 변하는 지점을 객체의 경계인 에지 부분으로 인식하여 출력하게 된다.

- 로버츠 크로스 탐지 방식(Roberts Cross) : 1963년에 L. Roberts가 제안하였으며, 2x2 크기의 필터(Filter)로 이미지의 대각선 방향의 변화율(기울기)을 구하여 에지를 검출한다[24]. 연산이 간단하고 빠르며, 대각선 방향의 에지를 잘 검출한다.

- 소벨 탐지 방식(Sobel) : 1968년에 I. Sobel과 G. Feldman이 제안하였으며, 3x3 소벨 필터로 픽셀값들의 변화율(기울기)을 구하여 에지를 검출한다[25]. 소벨 필터를 이미지의 x축과 y축 방향으로 각각 적용하여 구해진 기울기로부터 강도와 방향을 도출한 후, 이를 이용하여 에지를 판단한다.

- 캐니 탐지 방식(Canny) : 1986년에 J. Canny가 제안하였으며[26], 가장 유명하고 많이 사용된다. 노이즈 감소, 기울기 계산, 비최대 억제, 이중 임곗값, 히스테리시스(Hysteresis) 임곗값 처리 등의 단계별 프로세스(Process)를 거쳐 에지를 검출하며, 정확하고 선명한 에지를 탐지한다.

- 샤르 탐지 방식(Scharr) : 2000년에 H. Scharr가 제안하였으며, 소벨 탐지 방식과 원리는 같지만, 다른 필터 값을 적용하여 소벨 탐지 방식보다 정밀한 에지를 탐지할 수 있다[27].

III. The Proposed Scheme

본 연구는 선행 연구[16]의 후속 연구로서, 선행 연구에서 제안한 적대적 예제 탐지 모델이 다양한 알고리즘의 적대적 예제에 대해서도 효과적으로 탐지할 수 있음을 확인하고자 한다. 실험은 그림 1과 같이, 모델 학습, 4종류의 적대적 예제 생성, 클래스 예측 및 L1 거리 계산, 적대적 예제 탐지의 순으로 진행된다.

1. Model Training & Generating AEs

적대적 예제 탐지 테스트를 위한 선행 작업으로서, 모델 학습 및 적대적 예제 생성을 수행한다.

먼저, 기본 모델 학습을 위하여 이미지 분류에 주로 활용되는 CNN(Convolutional Neural Network) 모델 가운데 성능이 좋은 모델을 선정한 후, 준비된 클린 이미지 데이터 세트(Data Set)를 대상으로 학습을 수행한다. 기

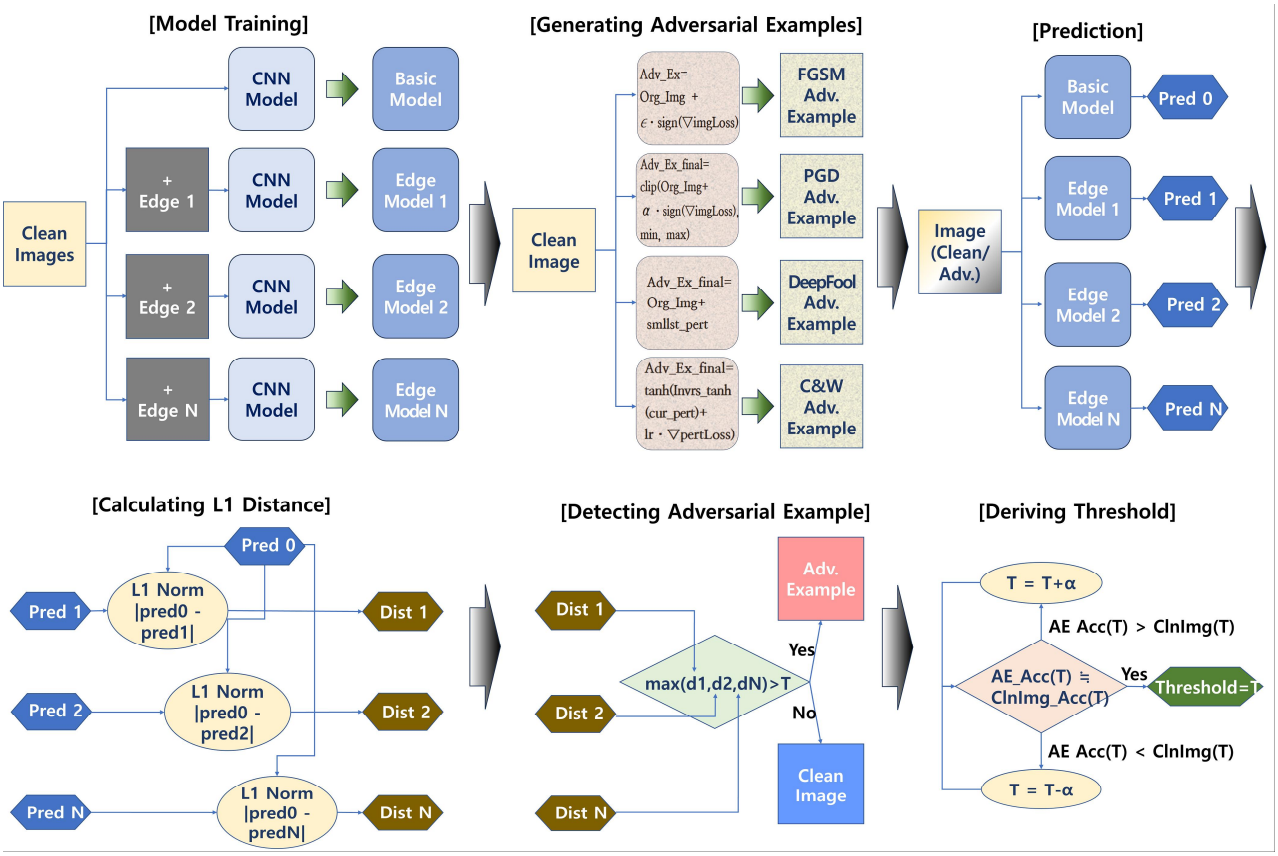

Fig. 1. Experiments Process

본 모델의 분류 정확도가 높을수록 적대적 예제 탐지 성능도 좋으므로, 최대한 높은 정확도를 얻을 수 있도록 학습을 수행해야 한다.

에지 모델 학습은 원본 이미지로부터 추출한 에지를 다시 해당 이미지에 결합하는 전처리(pre-processing)를 수행한 후, 분류 모델이 이를 학습하는 방식이다. 그림 2와 같이, 먼저 에지 탐지 기법을 사용하여 이미지(RGB 3채널)로부터 에지 이미지(1채널)를 추출한 후, 원본 이미지에 4번째 채널로 결합한다. 이렇게 전처리를 통하여 객체의 에지가 더해진 이미지를 분류 모델에 입력하여 학습을 수행한다.

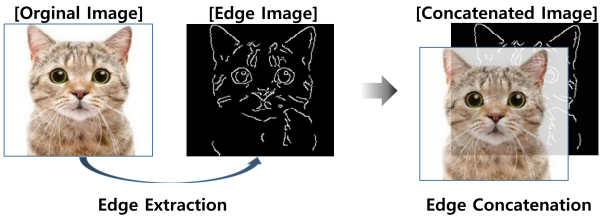

Fig. 2. Edge Extraction & Concatenation

본 탐지 모델에서는 복수 개의 에지 모델을 사용하므로, 각 에지 탐지 기법들을 사용하여 각각의 에지 모델들을 학습시킨다.

모델들의 학습이 완료되면, 테스트 데이터를 대상으로 FGSM, PGD, DeepFool, CW(L2) 알고리즘을 적용하여, 알고리즘별 적대적 예제들을 생성한다. 그리고 테스트 데이터(클린 이미지)와 알고리즘별로 생성된 적대적 예제들을 각각 섞어서, 총 4개의 알고리즘별 데이터 세트를 준비한다.

2. Calculating L1 Distance

적대적 예제 탐지의 첫 단계로서, 각 모델에 탐지 대상 이미지를 입력하여 예측값을 산출한다. 이때 예측값은 소프트맥스(Softmax) 함수를 적용하여 산출된 각 모델의 클래스별 확률값이다.

기본 모델과 에지 모델의 클래스별 확률값 간에 L1 Norm을 적용하여 두 모델 간 예측값의 거리(Distance)를 계산한다.

$$\text{L1 Distance} = \left(\sum_{i=1}^{n} \left|(\text{Prediction}(x)_i - \text{Prediction}^{\text{edged}}(x)_i)\right|\right)$$

L1 Norm은 벡터 요소들의 절댓값의 합으로 산출되는데, 본 실험에서는 위의 식과 같이 기본 모델의 클래스별 확률값(Prediction$(x)_i$)들과 에지 모델의 클래스별 확률값

(Prediction$^{edged}(x)_i$)들의 차가 벡터 요소들이 되며 해당 벡터값들의 절댓값의 합을 계산한다.

각 모델의 클래스별 확률의 최솟값은 0, 최댓값은 1, 모든 클래스별 확률의 합계는 1이기 때문에, 두 모델 간의 L1 Distance는 0~2 사이의 값을 갖는다. L1 Distance 값이 0에 가까울수록 클린 이미지일 가능성이 크고, 2에 가까울수록 적대적 예제일 가능성이 크다.

3. Detecting Adversarial Examples

전 단계에서 계산된 L1 Distance 값을 사용하여 입력 이미지에 대한 적대적 예제 여부를 판단한다. 먼저, 기본 모델과 복수 개의 에지 모델들 간의 L1 Distance 값들 가운데 가장 큰 값을 선택한다. 이것은 하나의 에지 모델이라도 입력 이미지가 적대적 예제일 가능성을 탐지한다면 해당 이미지를 적대적 예제로 판단하겠다는 탐지 정책을 반영한 것이다. 그런데, 분류 정확도가 낮은 에지 모델 중 하나가 클린 이미지를 오분류할 때도, 마찬가지로 해당 이미지를 적대적 예제로 판단(오탐)하게 될 수 있으므로, 본 탐지 모델에서는 기본 모델의 분류 정확도와 더불어 에지 모델의 분류 정확도가 매우 중요하다.

선택된 Max L1 Distance 값이 임계치(Threshold) 값보다 크면 적대적 예제로 분류한다. 여기서 임계치란 적대적 예제 여부를 판단하기 위한 임의의 기준값으로서, 임계치 값에 따라서 적대적 예제 탐지율과 클린 이미지를 적대적 예제로 판단하는 오탐률이 높아질 수도, 낮아질 수도 있으므로, 적대적 예제 탐지율을 높이는 동시에 오탐률을 최소화하는 값으로 설정해야 한다. 적대적 예제 탐지율과 오탐률은 Trade-off 관계에 있는데, 임계치 값을 작게 설정하면 적대적 예제 탐지율(진양성, True Positive)이 높아지지만 오탐률(위양성, False Positive)은 증가하게 된다. 반대로 임계치 값을 크게 설정하면 오탐률은 줄어들지만, 적대적 예제임에도 클린 이미지로 판단(위음성, False Negative)하게 되어 적대적 예제 탐지율이 낮아진다. 적대적 예제 탐지율을 유지하면서 오탐률을 줄이는 최선의 방법은 분류 모델들의 기본적인 분류 정확도를 향상하는 것인데[16], 본 연구에서는 선행 연구의 모델들보다 정확도가 향상된 모델들을 사용함으로써 탐지율 및 오탐률의 변화를 확인할 수 있었다.

IV. Experiments and Results

실험은 원본(클린) 이미지 데이터 세트를 사용하여 기본 모델 및 에지 모델을 학습시킨 후, 적대적 알고리즘별 적대적 예제 생성, L1 Distance 산출 및 임계치 설정, 적대적 예제 탐지 순으로 진행하였다.

1. Preparing Data Set

선행 연구와의 연속성을 고려하여, 데이터 세트는 선행 연구의 데이터 세트를 그대로 사용하였다. 즉, kaggle 사이트에서 30종의 동물 이미지[28]를 내려받은 후, 분류 모델과 실험 장비의 성능을 고려하여 선별된, 가로 또는 세로의 크기가 200 pixel 이상이면서 용량이 200kb 이하인 20,210개의 이미지이며, 훈련(Training), 검증(Validation), 평가(Test) 데이터를 7:2:1의 비율로 구분하여, 훈련 및 평가 과정에서 서로 섞이지 않도록 하였다.

2. Model Training & Evaluation

2.1 Model Training

본 연구는 기본 모델과 에지 모델 간의 분류 예측 결괏값의 차이를 통하여 적대적 예제를 탐지하고자 하므로, 각 모델의 기본적인 분류 정확도가 높지 않다면 의도치 않은 오탐이 발생하게 된다. 그러므로 오탐률을 줄이고 결과적으로 적대적 예제 탐지율을 높이기 위해서는 각 모델의 기본적인 분류 정확도를 높이는 작업이 매우 중요하다.

선행 연구에서는 EfficientNet-B1(이후 EB1) 모델로 실험을 위하여 준비한 데이터 세트를 사용하여 전이학습을 수행한 결과, 평가 데이터에 대하여 92.97%의 분류 정확도를 확인할 수 있었는데[16], 본 연구에서는 분류 정확도를 높이기 위하여 EfficientNet-B2(이후 EB2) 모델을 사용하여 전이학습을 수행한 결과, 96.88%의 분류 정확도를 확인하였다.

이미지 내 객체의 에지 탐지 방법도 선행 연구와 동일하게 Canny, Sobel, Scharr 탐지 방법을 적용하였으며, OpenCV 패키지에서 제공하는 에지 함수를 사용하였다. 적대적 예제, 특히 강한 노이즈가 적용된 적대적 예제에 대한 분류 정확도를 높이기 위하여 설정했던 에지 함수의 입력 파라미터도 같은 값으로 유지하였다. 그림 3은 원본 이미지에 대한 각 에지 탐지 방법에 따른 결과 이미지를 보여준다.

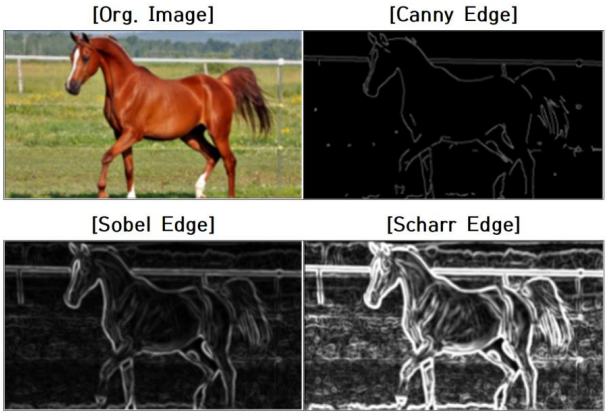

Fig. 3. Original Image and Edge Images

30종의 동물 이미지들에 대하여, 1채널로 추출된 에지 이미지를 3채널의 원본 이미지에 4번째 채널로 추가한 후 EB2 모델에 입력하여 전이학습을 수행하였다. 학습 종료 후 평가 데이터로 테스트한 결과, 기본 모델과 거의 같은 정확도를 보여주었던 선행 연구와 달리, 기본 모델보다는 조금씩 낮은 분류 정확도를 확인할 수 있었다. 표 2는 EB1 모델과 EB2 모델의 분류 정확도 비교 결과이다.

Table 2. Accuracies of Edge Models

Model	EfficientNet-B1	EfficientNet-B2
Basic Model	92.97%	96.88%
Canny Model	93.22%	95.05%
Sobel Model	92.28%	94.61%
Scharr Model	92.33%	93.57%

2.2 Adversarial Examples & Evaluation

다양한 적대적 예제 생성 알고리즘들에 대한 탐지 성능의 확인 및 탐지 결과의 비교를 위하여, FGSM을 포함하여 PGD, DeepFool, CW(L2) 알고리즘을 테스트 데이터에 적용하여 알고리즘별 적대적 예제를 생성하였다. 실제 구현은 Torchattacks 패키지에서 제공하는 함수를 사용하였으며, 노이즈 강도 등을 위한 파라미터 값의 설정은 함수 디폴트 값을 그대로 사용하였다.

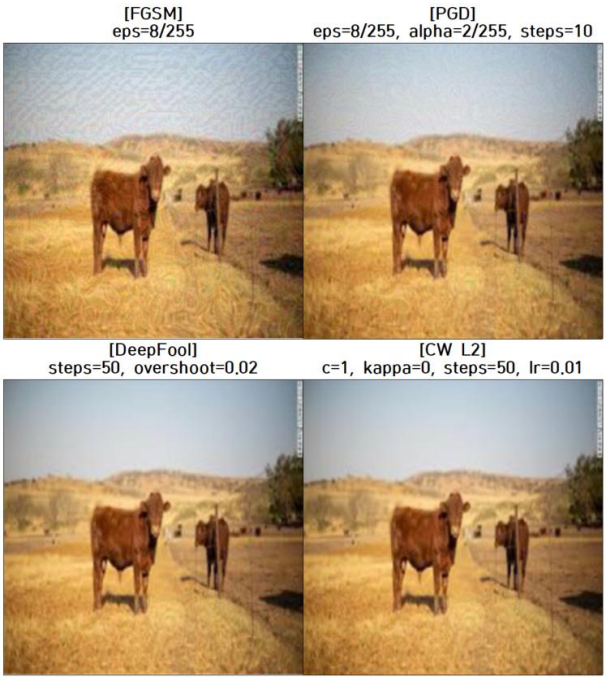

Fig. 4. Adversarial Images by Algorithm

그림 4(하늘 부분)를 보면 FGSM은 사람의 눈으로 노이즈의 확인이 가능하며, PGD도 약하기는 하지만 노이즈의 확인이 가능하다. 그러나 DeepFool과 CW(L2) 알고리즘이 적용된 적대적 예제들은 사람의 눈으로는 노이즈의 확인이 어렵기 때문에, 사람이 적대적 예제임을 알아차리지 못하도록 해야 하는 적대적 예제의 특성상 매우 강력한 알고리즘임을 알 수 있다.

알고리즘별 적대적 예제에 대하여 기본 모델과 에지 모델들의 분류 정확도를 평가하였다.

Table 3. Accuracies for Adversarial Algorithms

Adversarial Algorithm	Basic Model	Canny Model	Sobel Model	Scharr Model
FGSM	17.02%	62.74%	65.31%	76.30%
PGD	0.00%	83.23%	85.06%	88.77%
DeepFool	0.54%	92.78%	92.43%	92.38%
CW(L2)	1.63%	91.39%	91.69%	91.98%

표 3을 보면, FGSM 알고리즘의 적대적 예제에 대하여 기본 모델의 정확도는 17%로 많이 떨어졌지만 에지 모델들은 60~70% 대의 정확도를 유지하고 있으며, PGD, DeepFool, CW(L2) 알고리즘으로 생성된 적대적 예제들을 테스트한 결과, 기본 모델은 정답 클래스를 거의 예측하지 못했으나, 에지 모델들은 클린 이미지에 대한 분류 정확도와 큰 차이가 없음을 확인할 수 있다. 특히 강력한

적대적 예제 알고리즘으로 알려진 PGD, DeepFool, CW(L2) 알고리즘의 경우는 80~90%의 높은 분류 정확도를 보여주었다.

3. L1 Distance

3.1 Calculating L1 Distance

입력 이미지가 적대적 예제인가를 판단하기 위한 기본 모델과 에지 모델 간의 예측값의 차이(L1 Distance)는 두 모델의 클래스별 확률값들의 차에 L1 Norm을 적용하여 계산하였다.

Table 4. L1 Distances between Basic Model & Edge Models

Adversarial Algorithm	Label	Prediction			L1 Distance
		Basic Model	Edge Model		
Clean Image	15	15	Canny	15	0.000005
			Sobel	15	0.000020
			Scharr	15	0.000005
FGSM	1	9	Canny	9	0.019473
			Sobel	9	0.142475
			Scharr	9	0.688878
PGD	2	29	Canny	2	1.999996
			Sobel	2	1.999960
			Scharr	2	2.000000
DeepFool	25	13	Canny	25	1.362485
			Sobel	25	1.362432
			Scharr	25	1.362485
CW(L2)	5	17	Canny	16	1.984561
			Sobel	16	1.984049
			Scharr	16	1.984396

표 4는 클린 이미지와 알고리즘별 적대적 예제들에 대하여, 기본 모델과 에지 모델 간 L1 Distance를 계산한 결과 중 일부 예시이다.

클린 이미지에 대하여 기본 모델과 에지 모델들이 정답 클래스인 15를 예측했으며, 그 결과 L1 Distance들이 거의 0에 가까운 값임을 알 수 있다. 이 경우는 클린 이미지로 판단하게 된다.

FGSM 적대적 예제에 대하여 기본 모델과 에지 모델들이 모두 오답 클래스인 9를 예측했는데, 그 결과 산출된 L1 Distance 값이 클린 이미지보다는 크지만, 상당히 작은 값으로서, 결과적으로 적대적 예제 탐지에 실패하는 경우이다.

PGD 적대적 예제를 보면 기본 모델은 오답 클래스인 29를, 에지 모델들은 정답 클래스인 2를 예측했으며, 그 결과 L1 Distance의 값들이 2에 가까운 값임을 알 수 있다. 전형적인 적대적 예제 탐지에 성공하는 경우이다.

DeepFool 적대적 예제의 경우, PGD와 같이 기본 모델은 오답 클래스를 에지 모델들은 정답 클래스를 예측했는데, PGD와 달리 L1 Distance 값들이 1.4에 미치지 못하고 있다. 이 경우는 적대적 예제 판단 기준값(임계치)을 얼마로 설정하느냐에 따라 적대적 예제 탐지에 성공할 수도 있고 실패할 수도 있게 된다.

마지막으로 CW(L2) 적대적 예제는 기본 모델과 에지 모델들이 모두 정답 클래스인 5를 예측하지 못했는데, L1 Distance는 2에 가까운 값이 산출되었다. 본 논문의 적대적 예제 탐지 방법은 정답 클래스에 대한 예측 여부가 아니라, 모델 간 산출된 L1 Distance 값의 크기로 적대적 예제를 판단하는 것이기 때문에, 이 경우도 적대적 예제 탐지에 성공하게 된다.

3.2 Deriving threshold

산출된 L1 Distance 값으로 적대적 예제를 판단하기 위한 기준값인 임계치를 어떤 값으로 설정하느냐에 따라, 적대적 예제임에도 탐지에 실패하거나, 클린 이미지를 적대적 예제로 판단(오탐)하게 되므로, 임계치의 설정은 중요하다. 임계치를 작은 값으로 설정하면 적대적 예제 탐지율은 높아지지만, 클린 이미지에 대한 오탐률 또한 높아지게 된다. 반대로 임계치를 큰 값으로 설정하면 오탐률이 낮아지나 적대적 예제 탐지율 또한 낮아지게 된다.

본 실험에서는 임의의 값을 최초 임계치 값으로 설정한 후 적대적 예제 탐지 테스트(4. Detecting Adversarial Examples)를 진행하면서 적대적 예제 탐지율(진양성, True Positive)과 클린 이미지 정답율(진음성, True Negative)이 유사한 수치로 산출되었을 때의 기준값을 적대적 예제 탐지를 위한 최종 임계치 값으로 도출하였다. 임계치 값은 본 탐지 모델을 적용하는 실제 환경 및 상황에 따라 적절한 값으로 설정해야 한다.

Table 5. Appropriate Threshold for Each Algorithm

Adversarial Algorithm	Threshold	Adv. Image Correct Answer	Cln. Image Correct Answer
FGSM	1.23	91.06%	91.04%
PGD	1.85	95.20%	95.35%
DeepFool	1.35	91.99%	91.58%
CW(L2)	1.36	91.85%	91.92%

표 5를 보면 각 알고리즘에 따른 적정 임계치를 확인할 수 있는데, PGD 알고리즘의 임계치가 다른 알고리즘에 비해 큰 값으로 도출되었다. 이것은 표에서 확인할 수 있듯이 PGD 알고리즘의 적대적 예제에 대한 탐지율이 다른 알고리즘에 비해서 높기 때문(즉, 더 쉽게 탐지되기 때

문)이다. 임계치 값을 크게 설정하면 클린 이미지에 대한 정답율이 높아지는 대신 적대적 예제에 대한 탐지율이 낮아지는데, PGD 알고리즘으로 생성된 적대적 예제들은 상대적으로 L1 Distance 값이 크게 형성되었기 때문에, 임계치 값을 크게 설정하여 클린 이미지에 대한 정답율이 높아졌음에도 적대적 예제에 대한 탐지율이 낮아지지 않고 오히려 높아진 것이다.

4. Detecting Adversarial Examples

알고리즘별 적대적 예제 탐지 테스트를 수행하기 위해서, 클린 이미지(2,021)와 각 알고리즘으로 생성된 적대적 예제(각 2,021개)들을 섞어 4개의 데이터 세트를 준비한 후 알고리즘별 적정 임계치를 설정하여 탐지 테스트를 수행하였다. 선행 연구와 마찬가지로 일반 모델이 적대적 예제의 정답 클래스를 맞힌다면 해당 적대적 예제의 레이블을 클린 이미지로 변경하여 결과에 반영하였다(표 6에서 클린 이미지의 개수가 더 많은 이유)[16].

Table 6. Detection Results for Adversarial Examples of Each Algorithm

Adv. Algo. (Thrs.)	Adv. Imgs.		Cln. Imgs.		Prcsn. (TP/ TP+FP)	Recall (TP/ TP+FN)	F1-Scr. (PxRx2 /P+R)
	Corrct (TP)	Wrong (FN)	Correct (TN)	Wrong (FP)			
FGSM (1.23)	91.06%	8.94%	91.04	8.96	87.81%	91.06%	89.41%
	1,527	150	2,153	212			
PGD (1.85)	95.20%	4.80%	95.35%	4.65%	95.34%	95.20%	95.27%
	1,924	97	1,927	94			
Deep Fool (1.35)	91.99%	8.01%	91.58%	8.42%	91.53%	91.99%	91.76%
	1,849	161	1,861	171			
CW (1.36)	91.85%	8.15%	91.92%	8.08%	91.67%	91.85%	91.76%
	1,826	162	1,888	166			

표 6은 알고리즘별 적대적 예제에 대한 정답율(정탐율, 탐지율) 및 오답율, 클린 이미지에 대한 정답율 및 오답율(오탐률), 그리고 각 이미지 개수를 기반으로 산출된 정밀도(Precision)와 재현율(Recall), F1-Score를 보여주고 있다. 결과적으로, 에지를 이용한 적대적 예제 탐지 모델이, FGSM 알고리즘보다 더 강력한 적대적 예제 생성 알고리즘으로 알려진 PGD, DeepFool, CW(L2) 알고리즘에 대해 더 높은 탐지 성능을 가지고 있음을 확인할 수 있다.

추가로, 클린 이미지에 대한 오탐률을 줄이는 방안으로 선행 연구에서 제시했던[16], 기본적인 분류 정확도 향상 시 탐지율과 오탐률의 변화에 대한 실험을 수행하였다.

선행 연구와의 비교를 위하여, 같은 방법(클린 이미지와 각 epsilon 별 FGSM 적대적 예제들을 섞은 5개의 데이터 세트에 대하여 1.5의 임계치 값을 적용 등)으로 테스트를 수행하였다.

Table 7. Detection Results for FGSM Adversarial Examples (Threshold=1.5)

eps	Efficient Net	Adv. Image		Cln. Image	
		Correct	Wrong	Correct	Wrong
0.02	EB1	85.47%	14.53%	91.65%	8.35%
	EB2	89.94%	10.06%	93.01%	6.99%
0.05	EB1	84.64%	15.36%	90.74%	9.26%
	EB2	96.52%	3.48%	92.29%	7.71%
0.1	EB1	91.44%	8.56%	88.10%	11.90%
	EB2	99.64%	0.36%	90.00%	10.00%
0.2	EB1	95.47%	4.53%	88.02%	11.98%
	EB2	100,00%	0.00%	89.67%	10.33%
0.3	EB1	87.61%	12.39%	88.27%	11.73%
	EB2	99.95%	0.05%	89.53%	10.47%

표 7은 임계치 1.5를 기준으로 각 epsilon 별 FGSM 적대적 예제에 대한 탐지율 및 클린 이미지에 대한 정답률을 보여주고 있다. EB1(EfficientNet-B1)으로 표기된 행은 선행 연구의 실험 결과치이며, EB2(EfficientNet-B2)로 표기된 행들이 본 연구의 실험 결과치로서, EB2의 수치가 전체적으로 높음을 알 수 있다. 특히, 탐지율의 경우 선행 연구 대비 상당히 향상되었으며, 임계치를 조금 더 작게 설정했다면 클린 이미지에 대한 정답률 또한 더 높았을 것으로 추측된다. 즉, 분류 모델들의 기본적인 분류 정확도가 높아짐에 따라, 적대적 예제에 대한 탐지율이 향상되고 클린 이미지에 대한 오탐률이 낮아짐을 알 수 있다.

V. Conclusions

본 연구에서는, 기본 분류 모델과 에지 학습 모델 간의 L1 Distance 값으로 적대적 예제를 효과적으로 탐지할 수 있음을 증명한 선행 연구[16]를 기반으로, 더 강력한 알고리즘으로 생성된 적대적 예제들에 대한 탐지 성능을 확인하고자 하였다. 오탐률을 줄이기 위하여 선행 연구의 기본 모델인 EfficientNet-B1보다 더 높은 분류 정확도를 가진 EfficientNet-B2를 기본 모델로 하여 이를 기반으로 3종의 에지 모델들을 생성한 후, 30종의 동물 이미지를 대상으로 학습을 진행하였다. 이어서, FGSM 및 PGD, DeepFool, CW(L2) 등의 알고리즘을 적용한 적대적 예제를 생성하여 기본 모델과 에지 모델의 분류 정확

도를 측정한 결과, 클린 이미지에 대한 분류 정확도에 대비하여 기본 모델의 정확도는 FGSM : 17%, PGD/DeepFool/CW(L2) : 0~2%로 매우 낮아졌지만, 에지 모델들은 FGSM : 60~70%, PGD/DeepFool/CW(L2) : 80~90%의 높은 정확도를 유지하였다. 클린 이미지와 알고리즘별 적대적 예제로 구성된 4개의 데이터 세트를 대상으로 적대적 예제 탐지 모델의 성능을 테스트한 결과, 적대적 예제 탐지율은 91%~95%, F1-Score는 89~95%의 높은 탐지 성능을 보여주었으며, 특히 PGD, DeepFool, CW(L2) 등과 같이 강력한 적대적 알고리즘에서 오히려 더 높은 성능을 확인할 수 있었다. 그리고, 추가적인 실험을 통하여 클린 이미지에 대한 분류 정확도가 향상되었을 때 탐지율이 향상되고 오탐률이 낮아짐을 확인하였다.

정리하면, 객체의 에지를 학습한 분류 모델이 강력한 적대적 알고리즘에도 매우 강건함과 그 강건함을 기반으로 적대적 예제 탐지 모델이 다양한 종류의 적대적 예제들을 효과적으로 탐지해 내는 것을 확인할 수 있었다. 본 연구를 통하여, 적대적 예제에 취약하다고 알려진 딥러닝 모델의 강건성 향상 및 효과적인 적대적 예제 방어 수단을 제시하는 한편, 의료, 보안, 교통 등 다양한 산업에서 널리 활용되고 있는 딥러닝 분류 모델의 신뢰성을 제고할 수 있을 것이다.

한편, 본 연구에서는 단일 데이터 세트만을 사용하여 제안모델의 성능을 검증하였다. 후속 연구에서는 다양한 종류의 데이터 세트를 준비하여 제안모델이 동일한 탐지 성능을 보여줄 수 있는가에 대한 실험을 진행하고자 한다. 또한, 적대적 예제의 판단 기준인 임곗값의 설정과 관련하여, 적대적 공격 알고리즘을 미리 알 수 없는 현실적인 공격 상황을 고려하여, 보편적으로 적용할 수 있는 임곗값 도출에 관해서도 추가 연구를 진행하고자 한다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00217022).

REFERENCES

[1] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv Preprint arXiv:1412.6572, Dec. 2014, DOI: 10.48550/arXiv.1412.6572

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards deep learning models resistant to adversarial attacks," International Conference on Learning Representations, Feb. 2018, DOI: 10.48550/arXiv.1706.06083

[3] S. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574-2582, Jun. 2016, DOI: 10.1109/cvpr.2016.282

[4] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 IEEE Symposium on Security and Privacy (Sp), pp. 39-57, May 2017, DOI: 10.1109/sp.2017.49

[5] P. Chen, H. Zhang, Y. Sharma, J. Yi and C. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 15-26, 2017, DOI: 10.1145/3128572.3140448

[6] A. Athalye, N. Carlini and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in International Conference on Machine Learning, pp. 274-283, 2018, DOI: 10.48550/arXiv.1802.00420

[7] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno and D. Song, "Robust physical-world attacks on deep learning visual classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625-1634, Jun. 2018, DOI: 10.1109/cvpr.2018.00175

[8] A. Liu, J. Guo, J. Wang, S. Liang, R. Tao, W. Zhou, C. Liu, X. Liu and D. Tao, "X-adv: Physical adversarial object attacks against x-ray prohibited item detection," 32nd USENIX Security Symposium (USENIX Security 23), pp. 3781-3798, Aug. 2023, DOI: 10.48550/arXiv.2302.09491

[9] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," Pattern Recognition, vol. 110, pp. 107332, Feb. 2021, DOI: 10.1016/j.patcog.2020.107332

[10] E. Wong, L. Rice and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," arXiv Preprint arXiv:2001.03994, 2020, DOI: 10.48550/arXiv.2001.03994

[11] S. Gowal, C. Qin, J. Uesato, T. Mann and P. Kohli, "Uncovering the limits of adversarial training against norm-bounded adversarial examples," arXiv Preprint arXiv:2010.03593, 2020, DOI: 10.48550/arXiv.2010.03593

[12] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep

learning in computer vision: A survey," IEEE Access, vol. 6, pp. 14410-14430, 2018, DOI: 10.1109/access.2018.2807385

[13] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner and A. Madry, "Robustness may be at odds with accuracy," International Conference on Learning Representations, May 2019, DOI: 10.48550/arXiv.1805.12152

[14] A. Borji, "Shape defense," in I (Still) can'T Believe it's Not Better! Workshop at NeurIPS 2021, pp. 15-20, Feb. 2022, https://proceedings.mlr.press/v163/borji22a.html

[15] W. Xu, D. Evans and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," NDSS Symposium, 2018, DOI: 10.14722/ndss.2018.23198

[16] J. Shim and K. Jo, "Detecting Adversarial Examples Using Edge-based Classification," Journal of The Korea Society of Computer and Information, vol. 28, (10), pp. 67-76, 2023, DOI: 10.9708/jksci.2023.28.10.000

[17] A. N. Bhagoji, D. Cullina, C. Sitawarin and P. Mittal, "Enhancing robustness of machine learning systems via data transformations," in 2018 52nd Annual Conference on Information Sciences and Systems (CISS), pp. 1-5, 2018, DOI: 10.1109/CISS.2018.8362326

[18] N. Chattopadhyay, S. Chatterjee and A. Chattopadhyay, "Robustness against adversarial attacks using dimensionality," in International Conference on Security, Privacy, and Applied Cryptography Engineering, pp. 226-241, 2021, DOI: 10.1007/978-3-030-95085-9_12

[19] P. Samangouei, M. Kabkab and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," arXiv Preprint arXiv:1805.06605, 2018, DOI:10.48550/arXiv.1805.06605

[20] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille and K. He, "Feature denoising for improving adversarial robustness," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 501-509, 2019, DOI: 10.48550/arXiv.1812.03411

[21] C. Ferrari, F. Becattini, L. Galteri and A. D. Bimbo, "(Compress and Restore)N: A Robust Defense Against Adversarial Attacks on Image Classification," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, (1s), pp. 1-16, 2023, DOI: 10.1145/3524619

[22] D. Meng and H. Chen, "Magnet: A two-pronged defense against adversarial examples," in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135-147, Oct. 2017, DOI: 10.1145/3133956.3134057

[23] J. H. Metzen, T. Genewein, V. Fischer and B. Bischoff, "On detecting adversarial perturbations," International Conference on Learning Representations, 2017, DOI: 10.48550/arXiv.1702.04267

[24] L. G. Roberts, "Machine Perception of Three-Dimensional Solids," PhD Thesis. Massachusetts Institute of Technology, 1963, http://hdl.handle.net/1721.1/11589

[25] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," Pattern Classification and Scene Analysis, pp. 271-272, Jan. 1973, https://www.researchgate.net/publication/285159837_A_33_isotropic_gradient_operator_for_image_processing

[26] J. Canny, "A computational approach to edge detection," Readings in Computer Vision, pp. 184–203, 1987, DOI: 10.1016/b978-0-08-051581-6.50024-6

[27] H. Scharr, "Optimal operators in digital image processing," http://www.ub.uni-heidelberg.de/archiv/962, 2000.

[28] J. Bright, "ANIMALS(30 Animal Species for Easy train)," https://www.kaggle.com/datasets/jerrinbright/cheetahtigerwolf

Authors

Jaesung Shim received the M.S. degree in Information Security from Soongsil University, in 2008. He is currently a Ph.D. candidate in Computer Engineering at Chungbuk National University.

He is interested in computer vision and information security.

Kyuri Jo received her B.S. and Ph.D. degree in Computer Science and Engineering from Seoul National University in 2013 and 2018, respectively and worked as a postdoctoral researcher at Bio and Helath Informatics Lab., Seoul National University. She is currently an associate professor in the Department of Computer Engineering at Chungbuk National University since September 2019. Her current research interests include artificial intelligence, machine learning and bioinformatics.