Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.32, NO.5, Oct. 2022 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2022.32.5.975

이미지의 Symbolic Representation 기반 적대적 예제 탐지 방법*

박 소 희,[1†] 김 승 주,[2] 윤 하 연,[2] 최 대 선[3‡]
[1,2,3]숭실대학교 (대학원생, 학생, 교수)

Adversarial Example Detection Based on Symbolic Representation of Image*

Sohee Park,[1†] Seungjoo Kim,[2] Hayeon Yoon,[2] Daeseon Choi[3‡]
[1,2,3]Soongsil University (Graduate student, Undergraduate student, Professor)

요 약

딥러닝은 이미지 처리에 있어 우수한 성능을 보여주며 큰 주목을 받고 있지만, 입력 데이터에 대한 변조를 통해 모델이 오분류하게 만드는 적대적 공격에 매우 취약하다. 적대적 공격을 통해 생성된 적대적 예제는 사람이 식별하기 어려울 정도로 최소한으로 변조가 되며 이미지의 전체적인 시각적 특징은 변하지 않는다. 딥러닝 모델과 달리 사람은 이미지의 여러 특징을 기반으로 판단하기 때문에 적대적 예제에 속지 않는다. 본 논문은 이러한 점에 착안하여 이미지의 색상, 모양과 같은 시각적이고 상징적인 특징인 Symbolic Representation을 활용한 적대적 예제 탐지 방법을 제안한다. 입력 이미지에 대한 분류결과에 대응하는 Symbolic Representation과 입력 이미지로부터 추출한 Symbolic Representation을 비교하여 적대적 예제를 탐지한다. 다양한 방법으로 생성한 적대적 예제를 대상으로 탐지성능을 측정한 결과, 공격 목표 및 방법에 따라 상이하지만 specific target attack에 대하여 최대 99.02%의 탐지율을 보였다.

ABSTRACT

Deep learning is attracting great attention, showing excellent performance in image processing, but is vulnerable to adversarial attacks that cause the model to misclassify through perturbation on input data. Adversarial examples generated by adversarial attacks are minimally perturbated where it is difficult to identify, so visual features of the images are not generally changed. Unlikely deep learning models, people are not fooled by adversarial examples, because they classify the images based on such visual features of images. This paper proposes adversarial attack detection method using Symbolic Representation, which is a visual and symbolic features such as color, shape of the image. We detect a adversarial examples by comparing the converted Symbolic Representation from the classification results for the input image and Symbolic Representation extracted from the input images. As a result of measuring performance on adversarial examples by various attack method, detection rates differed depending on attack targets and methods, but was up to 99.02% for specific target attack.

Keywords: Image Classification, Adversarial Example, Symbolic Representation

Received(08. 19. 2022), Modified(09. 27. 2022), Accepted(09. 27. 2022)

* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발)과 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1A2C1014813)

† 주저자, sosohi@soongsil.ac.kr
‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

I. 서 론

인공지능은 빅데이터와 딥러닝 기술을 통해 이미지 처리 및 분류, 자연어 처리, 음성인식 등 다양한 분야에서 활용되고 있으며, 특히 이미지 처리에 있어 우수한 성능을 보여주면서 큰 주목을 받고 있다. 하지만, 인공지능의 핵심 기술인 딥러닝 또한 다양한 보안 취약점이 존재하며, 보안 위협에 대한 우려가 커지고 있다[1].

딥러닝 모델에 대한 보안 취약점에는 크게 학습 단계에서 발생하는 공격과 학습이 끝난 모델에 대해 활용 단계에서의 공격이 있다[2]. 먼저, 학습 단계에서의 공격에는 학습데이터에 악의적인 데이터를 주입하는 Poisoning Attack[3-4], 특정 트리거를 가진 데이터는 특정한 클래스로 오분류하도록 하는 Backdoor Attack[5-6]이 있으며, 활용단계에서의 공격에는 입력 데이터에 대한 최소한의 변조를 통해 모델이 오분류를 하게 만드는 적대적 공격(Adversarial Attack)[7-13], 입력 데이터와 출력값을 분석하여 모델을 추출하는 Model Extraction Attack[14], 모델의 학습데이터를 추출하는 Model Inversion Attack[15] 등이 있다.

특히 활용단계에서의 공격 중 적대적 공격의 경우 표지판 인식[16], 얼굴 인식[17], 의료영상 분석[18] 등 이미지 기반 딥러닝 모델에 대한 공격 연구가 활발히 이루어지고 있다. 그뿐만 아니라, 실제 교통 표지판에 스티커를 부착하여 움직이는 차량에서의 표지판 인식 모델이 84.8%로 오인식하게 하는 공격[16] 등 현실적으로 발생 가능한 상황을 가정한 물리적 공격(Physical Attack)에 관한 연구도 이뤄지고 있다[19-20]. 이에 따라, 적대적 공격에 대한 방어 방법 연구에 대한 필요성이 대두되고 있다.

적대적 공격을 통해 생성된 적대적 예제는 입력 이미지에 사람이 육안으로 식별하기 어려울 정도로 아주 작은 노이즈가 추가된다. 따라서 이미지의 전체적인 시각적 특징은 변하지 않으며 사람은 이미지의 여러 가지 특징을 기반으로 추론하여 판단하기 때문에 적대적 예제에도 속지 않는다.

본 논문은 이미지의 색상, 모양과 같은 시각적이고 상징적인 특징을 Symbolic Representation이라 정의하고 이를 활용하여 적대적 공격을 탐지하는 방법을 제안한다. 입력 이미지에 대한 분류모델 결과에 대응하는 Symbolic Representation과 입력 이미지로부터 추출된 Symbolic Representation을 비교하여 적대적 예제를 탐지한다. 입력 이미지가 정상이라면 딥러닝 모델이 올바르게 분류하기 때문에 분류결과로부터 변환된 Representation과 추출된 Representation이 동일하다. 하지만 적대적 예제는 모델이 오분류하도록 만들어진 데이터로 실제 클래스와는 다른 클래스로 분류되기 때문에 오분류된 결과로부터 변환된 Representation과 추출된 Representation은 다를 것이며, 이러한 점을 기반으로 적대적 예제를 탐지한다.

본 논문이 기여한 바는 다음과 같다.

- 이미지의 시각적이고 상징적인 특징을 Symbolic Representation이라 정의하며, 이미지의 Symbolic Representation을 활용한 적대적 예제 탐지모델을 제안하였다.
- 정상 이미지와 적대적 예제에 대하여 Symbolic Representation 추출 모델을 생성하였으며, 추출 정확도가 각각 99%, 94% 이상임을 보였다.
- 제안 모델을 통한 적대적 예제 탐지 결과, specific target attack에 대하여 최대 99%의 탐지율을 보였으며, Symbolic Representation 속성 수에 따른 실험을 통해 속성이 많을수록 탐지율이 더 향상하는 것을 보였다.

본 논문의 구성은 다음과 같다. 2장에서 적대적 공격 및 적대적 공격 방어 방법을 소개하고 3장에서 Symbolic Representation과 제안 방법인 Symbolic Representation 기반 적대적 예제 탐지 방법에 대해 설명한다. 4장에서 본 논문에서 제안한 방법을 평가하기 위해 다양한 적대적 공격을 통해 적대적 예제를 생성하고 탐지한 실험 결과를 도출하고, 5장과 6장에서 연구에 대한 고찰과 결론으로 논문을 맺는다.

II. 배경 및 관련 연구

2.1 적대적 공격

적대적 공격은 딥러닝 모델에 대한 대표적인 공격 방법의 하나로 입력 데이터에 대해 사람이 식별할 수 없을 정도의 최소한 변조를 통해 적대적 예제를 생성하여 모델의 오분류를 유도하는 공격이다. 적대적 공격은 공격 목표에 따라 Target Attack, Untarget Attack으로 나뉘는데 Target Attack은 적대적 예제가 특정한 클래스로 오분류하도록 하는 공격을 말하며, Untarget Attack은 단순하게 원본 클래스가 아닌 다른 클래스로 오분류하도록 하는 공격을 말한다[20].

적대적 공격은 Szegedy[7]에 의해 Box-Constrained L-BFGS 기반 공격이 처음으로 제안되었다. L-BFGS는 원본(정상) 이미지와 적대적 예제 사이의 유클리디안 거리를 최소화하면서 목표클래스로 오분류하는 적대적 예제를 생성하는 방법이다.

Goodfellow 등[8]은 한 번의 변조로 간단하고 빠르게 적대적 예제를 생성하는 FGSM(Fast Gradient Sign Method)를 제안하였다. 이는 분류모델의 손실함수를 통해 loss를 계산하고 기울기(Gradient)를 이용하여 원본 이미지를 변조할 방향과 변조 크기를 찾아 적대적 예제를 생성한다. 이 방법은 공격 속도는 빠르지만, 공격 성공률이 높지 않다는 한계가 있으며, 이를 개선하기 위해 Kurakin 등[9]은 FGSM을 여러 번 반복함으로써 조금씩 이미지를 변조하여 적대적 예제를 생성하는 BIM(Basic Iterative Method)을 제안하였다. 이는 각 반복에서 α만큼 이미지를 변조하며 clip 함수를 통해 최대 변조량 ε만큼 제한한다.

Madry 등[10]은 BIM을 일반화한 PGD(Projected Gradient Decent)를 제안하였으며, BIM은 변형이 없는 원본 이미지에서 공격을 시작하는 반면, PGD는 원본 이미지에 랜덤한 노이즈를 추가하여 공격을 시작한다. 두 공격은 공격 성공률은 높지만, 적대적 예제의 변조량의 크기가 크다는 한계가 있다.

Carlini와 Wagner[11]는 Distance Metric인 L0, L2, L∞를 기반으로 적대적 예제를 생성하는 3가지 공격을 제안하였으며, 가장 많이 사용되는 공격인 L2-공격은 다음과 같이 정의된다.

$$\text{minimize} \ \| x - x' \|_2 + c \cdot f(x') \tag{1}$$

$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -k) \tag{2}$$

여기서 $Z(x)$는 딥러닝 모델의 logit vector를 나타내며, k는 confidence를 나타낸다. k는 딥러닝 모델이 공격을 통해 생성된 적대적 예제를 공격 목표 클래스(target)로 분류하는 confidence를 말하며 이 값이 커질수록 공격 성공률은 높아지지만, 원본 이미지와 적대적 예제의 거리 차이가 커진다.

그 외에도 DeepFool[12], MIM[13] 와 같은 다양한 적대적 공격들이 존재하며, 기존의 적대적 공격을 방어하는 기술이 나오기 시작하면서 이러한 방어기술에 대응하는 심화된 공격 기술들 또한 제시되고 있다.

2.2 적대적 공격 탐지 방법

적대적 예제를 탐지하기 위해 일반적으로 정상데이터와 적대적 예제의 특성 차이를 비교한다.

Meng 등[21]은 입력 이미지와 오토인코더(Autoencoder)를 활용하여 복원된 입력 이미지의 차이를 통해 적대적 예제를 탐지하고 디노이징까지 하는 Magnet을 제안하였다. 먼저, 정상 이미지와 오토인코더를 통해 디노이징된 이미지가 최소화되도록 오토인코더를 학습시킨다. 이후 활용단계에서 입력 이미지와 학습된 오토인코더를 통해 디노이징된 이미지의 차이가 임계값보다 크면 적대적 예제로 탐지한다. 정상으로 판단된 이미지는 오토인코더를 통해 재복원하여 딥러닝 모델에 입력하며 이때, 탐지 못 한 적대적 예제가 있더라도 오토인코더를 통해 노이즈의 효과를 감소시키게 된다.

Xu 등[22]은 입력 이미지를 변조하여 변조된 이미지와 원본 이미지에 대한 모델의 출력값 차이를 통해 적대적 예제를 탐지하는 Feature Squeezing을 제안하였다. 이미지 변조는 공격자의 공격 공간을 줄이기 위해 8-bit로 표현되는 이미지의 color-bit를 축소하는 방법과 노이즈 영향을 감소시키기 위해 smoothing 방법을 적용한다. 이 과정을 통해 변조된 입력 이미지의 모델 출력값과 변조되지 않은 원본 입력 이미지에 대한 모델 출력값의 차이를 계산하여 차이가 임계값 보다 크면 적대적 예제로 탐지한다.

Freitas[23]는 입력 이미지에 대해 robust feature를 추출하고 입력 이미지에 대한 분류결과로부터 기대되는 feature와 비교함으로써 적대적 예제를 탐지하는 Unmask를 제안하였다. 이때, feature가 같으면 정상으로 판단하며, 다르면 적대적 예제로 탐지하고 탐지된 적대적 예제의 robust feature를 기반으로 가장 높은 유사도를 가진 클래스로 입력 이미지를 재분류하는 과정을 수행한다.

이 외에도 입력 데이터의 노이즈를 완화 및 제거하는 디노이징 방법 및 강건한 딥러닝 모델을 구축하기 위한 적대적 학습방법 등 다양한 방어 방법이 있다[2,20].

III. Symbolic Representation 기반 적대적 예제 탐지 방법

본 논문은 적대적 예제는 사람이 식별하기 어려울 만큼의 최소한으로 변조되어 오히려 사람은 이미지의 전체적인 특징을 기반으로 적대적 예제에 속지 않는다

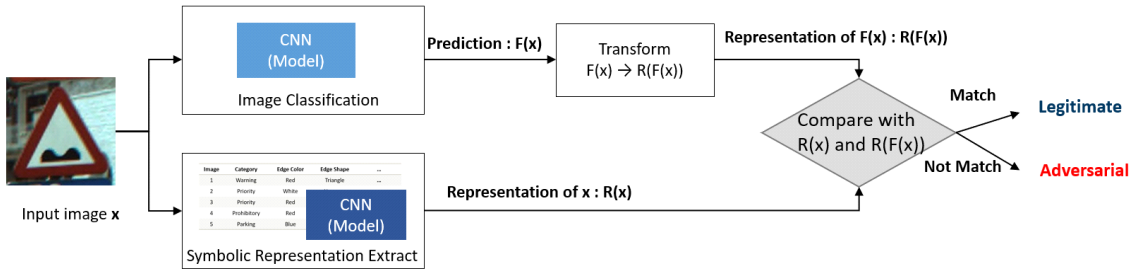

Fig. 1. Adversarial Attack Detection Based on Symbolic Representation

는 점에서 착안하여, 이미지의 전체적인 시각적 특징을 Symbolic Representation이라 하며, 이를 활용한 적대적 예제를 탐지하는 방법을 제안한다.

3.1 Symbolic Representation 및 추출 모델 생성

이미지 객체의 모양, 색깔과 같이 이미지의 시각적이고 상징적인 특징을 Symbolic Representation이라 정의하고, 이미지의 속성별 정보를 추출하여 생성한다. Symbolic Representation은 각 속성(ex. Color)에 대한 값(ex. Red)으로 구성되며, 이미지 클래스별 Symbolic Representation을 사전에 정의한 후, 이미지와 Representation을 기반으로 CNN(Convolution Neural Network)을 학습하여 Symbolic Representation 추출 모델을 생성한다.

Symbolic Representation 추출 모델은 각 속성에 대한 값을 추출하며, {X: 이미지, Y: Symbolic Representation}′을 학습한다. Y는 이미지별 해당하는 Representation에 대하여 '1'의 값을 가지며 그렇지 않으면 '0'의 값을 가지는 이진화된 값으로 구성된다.

3.2 Symbolic Representation 기반 적대적 예제 탐지

Symbolic Representation 기반 적대적 예제 탐지과정은 Fig. 1. 과 같다. 먼저, 딥러닝 모델(분류기)에 이미지가 입력되면 분류를 수행하게 되고, 분류결과에 대응하는 사전에 정의된 Symbolic Representation으로 변환한다. 그리고 Symbolic Representation 추출 모델을 통해 입력 이미지에 대한 Symbolic Representation을 추출한다. 최종적으로 변환된 Representation 값과 추출된 Representation 값을 비교하여 같으면 정상, 그렇지 않으면 비정상(적대적 예제)이라고 판단한다.

정상 이미지라면 딥러닝 모델이 올바르게 분류하기 때문에 변환된 Representation과 이미지로부터 추출된 Representation이 동일한 값을 가질 것이다. 하지만 적대적 예제는 모델이 오분류하도록 만들어진 데이터로 분류결과가 실제 클래스와는 다른 클래스를 도출하게 된다. 따라서, 오분류된 결과로부터 변환된 Representation과 이미지로부터 추출된 Representation은 다른 값을 가지게 될 것이다.

IV. 실험 및 실험 결과

본 논문에서 제안하는 적대적 예제 탐지 방법은 입력 이미지에 대한 분류결과에 대응하는 Symbolic Representation과 입력 이미지로부터 추출한 Symbolic Representation을 비교함으로써 적대적 예제를 탐지한다. 제안한 적대적 예제 탐지 방법을 평가하기 위해 분류모델과 Symbolic Representation 추출 모델을 생성한다. 그리고 정상 이미지와 적대적 공격방법인 BIM(Basic Iterative Method)[9], PGD(Projected Gradient Decent)[10], CW(Carlini&Wagner) Attack[11]을 통해 생성한 적대적 예제를 대상으로 탐지성능을 측정하였다.

4.1 실험 환경 설정

본 논문에서는 벨기에의 교통표지판 분류 데이터 세트(BTSCD, Belgium Traffic Sign Classification Dataset)[24]를 사용하였으며, 교통표지판 분류모델을 대상으로 적대적 예제 생성 및 탐지를 수행하였다.

벨기에의 교통표지판 분류 데이터 세트는 약 63종류의 교통표지판으로 구성되어있으며, 학습데이터 세트는 4,575개로 최소 22x20에서 최대 724x529 사

Table 1. 3-Attribute based Symbolic Representation

Attribute	Value
Color	Red, Redblue, Blue, Yellow
Shape	Circle, Diamond, Hexagon, InverseTriangle, Triangle, Square, RectanlesUp, RectanglesDown
Figure	Blank, Character, Direction, Number, Picture

Table 2. 6-Attribute based Symbolic Representation

Attribute	Value
Category	Warning, Parking, Prioirity, Prohibitory, Mandatory, Indicatory
Edge Color	Red, Blue, White
Edge Shape	Circle, Diamond, Hexagon, InverseTriangle, Triangle, Square, RectanlesUp, RectanglesDown
Central Color	Red, Blue, White, Yellow
Figure	Blank, Character, Direction, Number, Picture
Diagonal Line	Zero, One, Two

이즈의 이미지로 구성되어있으며, 테스트 데이터 세트는 2,520개의 이미지로 구성되어있다.

교통표지판 분류모델을 생성하기 위해 CNN(Convolution Neural Network)을 사용하였으며, 제공하는 학습데이터 세트를 통해 train accuracy가 100%, validation accuracy가 98.6%로 학습하였으며 결과적으로, test accuracy가 96.9% 정확도를 가진 분류모델을 생성하였다. 해당 교통표지판 모델을 공격대상으로 적대적 예제를 생성하고 제안한 탐지 방법을 적용하여 실험을 진행하였다.

4.2 Symbolic Representation 추출 모델 생성

데이터 세트의 클래스별로 Symbolic Representation 속성 및 값을 정의하여 Symbolic Representation 추출 모델을 생성하였다. 그리고 속성 개수에 따른 결과를 비교하기 위해 속성을 3개, 6개로 설정하여 Representation을 정의하였으며, Table 1.과 Table 2.과 같다. 3-Attribute는 이미지의 색(color), 모양(shape), 그림(figure)으로 정의하였으며 6-Attribute는 3-Attribute를 보다 세분화하여 유형(category), 가장자리 색(edge color), 가장자리 모양(edge shape), 중앙색(central color), 그림(figure), 대각선(diagonal line)으로 정의하였다.

앞서 정의한 Symbolic Representation 개수에 따라 3-Attribute에 기반한 Representation 추출 모델과 6-Attribute에 기반한 Representation 추출 모델, 두 가지를 생성하였다.

Symbolic Representation 추출 모델은 CNN을 기반으로 이미지와 이미지 클래스별 Symbolic Representation 학습을 통해 생성하였으며, Representation 속성값을 추출하기 위해 Multi-label(Binary Relevance)로 학습하였다. 즉, 이미지 x의 Symbolic Representation이 'Red', 'Triangle', 'Number'라면 x에 대한 y 값은 전체 속성값 {Y: Red, Blue, Yellow, Circle, Diamond, , Triangle, …}에 대하여 {1,0,0,0,0,1,…}의 값을 가진다.

Symbolic Representation 추출 모델을 생성한 결과, 3-Attribute에 기반한 Representation 추출 모델의 정확도는 99.37%를 보였으며, 6-Attribute에 기반한 Representation 추출 모델의 정확도는 98.93%를 보였다.

4.3 적대적 예제 생성

적대적 공격방법인 BIM(Basic Iterative Method), PGD(Projected Gradient Decent), CW(Carlini&Wagner) Attack을 사용하여 교통표지판 분류모델을 공격대상으로 오분류를 위한 적대적 예제를 생성하였다. 앞서 클래스별로 3개, 6개의 속성을 가지는 Symbolic Representation을 정의하였지만, 다른 클래스더라도 이미지의 유사성으로 인해 Symbolic Representation이 동일한 경우가 있으며, 이를 고려하여 다음과 같이 공격 목표를 3개로 설정하여 적대적 예제를 생성하였다.

1) Specific Target Attack: Symbolic Representation이 다른 클래스를 Target으로 공격
2) Random Target Attack : 무작위 클래스를 Target으로 공격
3) Untarget Attack : Target 없이 원본 클래스가 아닌 다른 클래스로 공격

공격알고리즘별 파라미터는 BIM, PGD의 target attack(specific, random)의 경우 ε=0.3, 0.251, untarget attack은 ε=0.063으로 설정하였으며, CW Attack의 경우 binary search step=13, k=0으로 설정하였다. BTSCD의 테스트 데이터 세트를 사용하여 클래스마다 20개의 적대적 예제를 생성하였으며, 테스트 데이터 세트의 클래스별 데이터 수가 20개 미만인 클래스가 존재하며 해당 클래스는 20개 미만의 적대적 예제가 생성되었다.

각 공격 알고리즘별 공격 성공률은 Table 3.과 같다. specific target attack의 경우 Symbolic Representation이 다른 클래스를 target으로 공격하였기 때문에 3-Attribute based Symbolic Representation을 기준으로 Representation이 다른 클래스로 공격한 결과와 6-Attribute based Symbolic Representation을 기준으로 Representation이 다른 클래스로 공격한 결과로 나뉜다.

교통표지판 분류모델에 대한 공격 성공률은 공격 목표 및 방법에 따라 상이하지만, 최소 85%에서 최대 97%를 보였으며, 공격이 성공한 적대적 예제만을 탐지성능 측정을 위해 사용하였다. Fig. 2.는 CW(specific target attack)를 사용하여 생성한 적대적 예제를 보여주며 왼쪽은 원본 이미지, 중간은 공격을 위한 target label의 이미지, 오른쪽은 원본 이미지를 target label로 오분류하도록 생성된 적대적 예제이다. 생성된 적대적 예제는 시각적으로 원본 이미지와 큰 차이가 없음을 알 수 있다.

4.4 실험 결과

본 논문에서 제안한 적대적 예제 탐지 방법을 평가하기 위해 정상 이미지 및 적대적 예제에 대하여 성능을 측정한다.

제안 방법은 입력 이미지가 적대적 예제인지 아닌지 판단하기 때문에 정상 이미지에 대한 오탐이 얼마나 발생하는지 평가할 필요가 있다. 특히 적대적 예제를 올바르게 탐지할지라도 정상 이미지에 대한 오탐율이 높아질 수 있으며, 좋은 탐지모델은 탐지율은 높은 동시에 오탐율은 낮아야 한다. 따라서, 정상 이미지(Negative)를 정상으로 올바르게 판단하였는지 평가하는 True Negative Rate(정탐율)와 정상 이미지를 적대적 예제(Positive)라고 판단하였는지 평가하는 False Positive Rate(오탐율)를 측정하며 식은 (3),(4)와 같다.

또한, 제안한 적대적 예제 탐지 방법은 Symbolic

Table 3. Adversarial Attack Success Rate

Attack	BIM	PGD	CW
Specific Target (3-Attribute)	89.46%	94.63%	92.16%
Specific Target (6-Attribute)	84.46%	95.19%	93.26%
Random Target	85.97%	94.36%	92.98%
Untarget	97.52%	97.52%	96.56%

Representation을 활용하여 적대적 예제를 탐지하기 때문에 이미지에 대한 Symbolic Representation 추출 모델의 정확도가 낮다면 탐지 결과를 신뢰할 수 없다. Symbolic Representation 추출 모델은 정상 이미지 뿐 아니라 적대적 예제에 대해 원본 이미지(변조 전)와 같이 Symbolic Representation을 잘 추출하여야 한다. 예를 들어, 원본 이미지의 label이 0이고 해당 이미지를 label 1로 오분류하도록 만들어진 적대적 예제가 있을 때, 적대적 예제로부터 Symbolic Representation을 추출하면 label 0을 가진 원본 이미지(변조 전)의 Symbolic Representation이 나와야 한다. 따라서, 적대적 예제에 대해 Symbolic Representation을 정확하게 추출하는지 평가하기 위해 Exact Match Rate(정확도)를 측정하며 식은 (5)와 같다.

Fig. 2. Example of Adversarial Attack(CW) (left:origin image, central:target image, right:adversarial example)

또한, 제안 방법이 적대적 예제를 올바르게 탐지하였는지 평가하기 위해 True Positive Rate(Detection Rate, 탐지율)를 측정하며 식은 (6)과 같다.

$$\mathrm{TNR} = \frac{True\ Nagative}{True\ Nagative + False\ Positive} \quad (3)$$

$$\mathrm{FPR} = \frac{False\ Positive}{True\ Nagative + False\ Positive} \quad (4)$$

$$\mathrm{EMR} = \frac{The\ number\ of\ Correct\ Predictions}{Total\ \ Examples} \quad (5)$$

$$\mathrm{TPR} = \frac{True\ Positive}{True\ Positive + False\ Nagative} \quad (6)$$

4.4.1 정상 이미지에 대한 탐지성능 평가

앞서 4.2에서 Symbolic Representation에 대하여 속성을 3개, 6개로 설정하고 3-Attribute based Representation 추출 모델 및 6-Attribute based Representation 추출 모델을 생성하였다.

정상 이미지에 대하여 Symbolic Representation 추출 모델 기반으로 탐지성능을 측정하였으며, True Negative, False Positive Rate에 대한 결과는 Table 4.와 같다. 3-Attribute 기반 Representation을 활용한 탐지 방법의 경우 정상 이미지가 입력되었을 때, 98.29%로 정상으로 올바르게 판단하였으며, 1.71%로 적대적 예제로 잘못 탐지하였다. 6-Attribute 기반 Representation을 활용한 경우, 97.71%로 정상으로 올바르게 판단하였으며 2.30%로 적대적 예제라고 잘못 판단하였다.

4.4.2 적대적 예제에 대한 Symbolic Representation 추출 성능평가

각 공격 목표(specific target, random target, untarget)에 따라 BIM, PGD, CW Attack을 통해 생성된 적대적 예제에 대한 Symbolic Representation 추출 모델의 성능을 평가하였으며 결과는 Table 5.와 같다. Table 5.는 Attribute 수에 따라 3-Attribute 기반 Symbolic Representation 추출 정확도 및 6-Attribute 기반 Symbolic Representation 추출 정확도를 나타낸다.

Table 4. Evaluation of Symbolic Representation based Detection for Clean Image

Type of Symbolic Representation	True Negative Rate	False Positive Rate
3-Attribute based Representation	98.29%	1.71%
6-Attribute based Representation	97.70%	2.30%

Table 5. Evaluation of Symbolic Representation Extraction for Adversarial Example

Adversarial Attack		3-Attribute based	6-Attribute based
		Exact Match Rate	
Specific Target Attack	BIM	81.22%	85.50%
	PGD	88.48%	92.34%
	CW	99.40%	98.53%
Mean		89.70%	92.12%
Random Target Attack	BIM	83.39%	85.12%
	PGD	87.76%	92.42%
	CW	99.26%	98.67%
Mean		90.14%	92.07%
Untarget Attack	BIM	99.15%	98.31%
	PGD	98.87%	98.37%
	CW	99.72%	98.87%
Mean		99.25%	98.52%

3-Attribute 기반 Symbolic Representation 추출 모델의 경우 specific target attack으로 생성된 적대적 예제에 대하여 평균 89.70%의 Representation 추출 정확도를 보였으며, random target attack으로 생성된 적대적 예제는 평균 90.14%, untarget attack으로 생성된 적대적 예제는 평균 99.25%의 정확도를 보였다. 그리고 6-Attribute 기반 Symbolic Representation 추출 모델은 specific target attack으로 생성된 적대적 예제에 대하여 평균 92.12%의 Representation 추출 정확도를 보였으며, random target attack으로 생성된 적대적 예제는 평균 92.07%, untarget attack으로 생성된 적대적 예제는 평균 98.52%의 정확도를 보였다. 즉, 적대적 예제에 대해 평균 94% 이상 Representation을 올바르게 추출함을 보였다.

앞서, 정상 이미지에 대한 추출 정확도 99.37%(3-Attribute based), 98.93%(6-Attribute ba

Table 6. Evaluation of Proposed Method for Adversarial Example

Adversarial Attack		3-Attribute based		6-Attribute based	
		Detection Rate	Ratio of Attacking Same Representation	Detection Rate	Ratio of Attacking Same Representation
Specific Target Attack	BIM	98.02%	0%	98.21%	0%
	PGD	98.98%	0%	99.13%	0%
	CW	99.85%	0%	99.74%	0%
Random Target Attack	BIM	85.92%	12.48%	89.12%	9.12%
	PGD	87.61%	11.66%	90.52%	8.60%
	CW	88.46%	11.54%	91.42%	8.43%
Untarget Attack	BIM	48.12%	51.46%	63.75%	34.59%
	PGD	48.80%	50.63%	65.16%	33.71%
	CW	49.15%	50.85%	64.08%	35.35%

sed)에 비해 비교적 저조한데, 이는 적대적 예제의 경우 공격대상 모델의 오분류를 유도하기 위해 노이즈가 추가되어 변조된 이미지이기 때문에 변조량이 증가할수록 추출 정확도는 떨어지게 된다. 또한, untarget attack으로 생성된 적대적 예제의 추출 정확도가 target Attack으로 생성된 적대적 예제에 비해 높게 나타나는 것은 untarget attack이 변조량이 더 적기 때문이다.

4.4.3 적대적 예제에 대한 탐지성능 평가

Symbolic Representation 기반 적대적 예제 탐지 방법의 성능을 평가하기 위해 각 공격 목표(specific target, random target, untarget)에 따라 BIM, PGD, CW Attack을 통해 생성된 적대적 예제를 대상으로 탐지성능을 평가하였다.

Table 6.은 3-Attribute 및 6-Attribute 기반 Symbolic Representation을 활용한 적대적 예제 탐지성능을 공격 목표별로 비교한다. 표의 Detection Rate는 적대적 예제에 대한 탐지한 적대적 예제의 비를 말하며, Ratio of Attacking Same Representation은 입력 이미지에 대하여 적대적 공격을 통해 생성된 적대적 예제 중 입력 이미지와 동일한 Representation을 가진 이미지로 공격한 비를 말한다.

공격 목표에 따라 비교하였을 때, specific target attack에 대한 탐지율은 약 98~99% 이상으로 대부분의 적대적 예제를 탐지하며 높은 탐지율을 보였고 random target attack은 약 85~91%의 탐지율을 보였다. 반면에, untarget attack에 대한 탐지율은 48~65%로 낮게 나타났다. 이는 적대적 공격을 할 때, Symbolic Representation이 동일한 이미지로 공격한 경우를 탐지하지 못하였기 때문이다.

앞서 언급한 바와 같이 Symbolic Representation을 정의할 때, 일부 클래스의 경우 이미지의 유사성으로 인해 동일한 값을 가진다. 따라서, Symbolic Representation이 다른 클래스를 target으로 공격한 specific target attack 외의 random target attack과 untarget attack의 경우 동일한 Representation을 가진 클래스로 공격하는 경우가 발생한다. 본 논문에서 제안하는 방법은 이미지의 Symbolic Representation 비교함으로써 공격을 탐지하기 때문에, 공격 전 원래 클래스와 동일한 Representation을 가진 이미지로 오분류 되도록 공격을 하면 탐지하지 못한다. Table 6.에서 각 공격 별로 동일한 Representation을 가진 이미지로 공격한 비를 보면 random target attack과 untarget attack의 경우 그 비율이 증가함에 따라 탐지율이 감소하는 것을 알 수가 있다. 또한, 동일한 Representation을 가진 이미지로 공격함에 따라 탐지하지 못하는 경우를 제외하면, 대부분의 적대적 예제를 탐지하는 것을 알 수 있다.

Table 7. Comparison of Detection Rate by number of Symbolic Representation for Adversarial Example

Adversarial Attack	3-Attribute based	6-Attribute based	Rate of Increase
Specific Target Attack	98.95%	99.03%	0.08%
Random Target Attack	87.33%	90.35%	3.02%
Untarget Attack	48.69%	64.33%	15.46%

또한, 속성 개수에 따른 탐지율을 비교하였으며, 결과는 Table 7.과 같다. 3-Attribute에 기반한 탐지 방법의 경우 specific target attack에 대해 평균 약 98.95%, random target attack에 대해 87.33% 그리고 untarget attack에 대해 48.69%의 탐지율을 보였으며 6-Attribute에 기반한 탐지 방법의 경우 specific target attack에 대해 평균 약 99.03%, random target attack에 대해 90.35% 그리고 untarget attack에 대해 64.33%의 탐지율을 보였다. 공격목표별로 속성 개수에 따라 탐지율을 비교한 결과, 6-Attribute에 기반한 탐지 방법이 3-Attribute에 기반한 탐지 방법 보다 약 0.08%, 3.02%, 15.46% 높은 탐지율을 보였으며, Symbolic Representation의 속성 개수가 많을수록 탐지율이 높다는 것을 알 수 있다.

V. 고 찰

본 논문은 이미지의 상징적인 특징인 Symbolic Representation 기반 적대적 예제 탐지 방법을 제안하였다. 이미지의 모양, 색깔과 같이 시각적이고 상징적인 특징을 Symbolic Representation이라 하고, 속성과 그에 따른 속성값으로 구성하였다. 실험에서는 속성의 개수를 3개, 6개로 설정하여 6개일 때의 탐지성능이 더 높은 결과를 보였다.

하지만 속성 수가 각 이미지를 구분할 만큼 충분하지 않은 경우, 유사한 이미지는 속성에 대한 값이 모두 같아진다. 즉, 데이터 세트 내의 다른 클래스의 이미지더라도 같은 Symbolic Representation을 가지는 경우가 발생한다. 본 논문에서 제안하는 방법은 Symbolic Representation을 비교하여 탐지하기 때문에 같은 값을 가지는 이미지로 공격할 경우 탐지하기 어렵다는 한계가 있다. 따라서, 공격 탐지율을 향상시키기 위해 Symbolic Representation을 보다 세분화하는 것이 필요하다. 다만, 본 논문의 실험 결과를 보면 Symbolic Representation을 세분화할수록 Symbolic Representation 추출 정확도가 비교적 낮아질 수 있으며 이러한 점을 고려할 필요가 있다.

또한, 제안하는 탐지 방법은 분류모델 외의 CNN을 사용하여 생성한 Symbolic Representation 추출 모델이 필요하다. 이때, 추출 모델도 딥러닝을 활용하기 때문에 공격자는 추출 모델까지 속이기 위한 적대적 공격을 시도할 수 있다. 따라서, 딥러닝이 아닌 기존의 패턴인식 분야에서 사용되는 색상 정보, 형태 정보 등 추출하는 방법을 활용하여 Representation을 추출한다면 보다 적대적 공격에 강건한 모델을 만들 수 있다.

VI. 결 론

본 논문은 이미지의 시각적이고 상징적인 특징으로 Symbolic Representation을 활용하여 적대적 예제를 탐지하는 방법을 제안하였다. 제안 방법을 평가한 결과, 3-Attribute 기반한 탐지 방법의 경우 specific target attack에 대해 98.95%, random target attack에 대해 87.33%, untarget attack에 대해 48.69%의 탐지율을 보였다. 속성을 확장한 6-Attribute 기반한 탐지 방법의 경우 specific target attack에 대해 99.03%, random target attack에 대해 90.35%, untarget attack에 대해 64.33%의 탐지율을 보였다. 또한, 속성 개수에 따른 탐지율을 비교하였을 때, 3-Attribute보다 속성이 많은 6-Attribute에 기반한 탐지 방법이 탐지율이 높게 나타나는 결과를 보였다.

향후에는 Symbolic Representation을 효과적으로 추출하는 방법에 대한 연구와 본 논문에서 제안하는 방법은 적대적 예제를 탐지하기 위해 Symbolic Representation을 추출하는 별도의 네트워크가 존재하고 이에 대한 공격을 피하기 위한 연구를 진행할 예정이다.

References

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

[2] M. Xue, C. Yuan, H. Wu, Y. Zhang and W. Liu, "Machine learning security: Threats, countermeasures, and evaluations.", IEEE Access, Vol. 8, pp. 74720-74742, April, 2020.

[3] C. Yang, Q. Wu, H. Li, and Y. Chen, "Generative poisoning attack methodagainst neural networks," arXiv preprint arXiv:1703.01340. March, 2017

[4] L. Munoz-Gonzalez, B. Biggio, A. Demontis, A. Paudice,V. Wongrassamee,

E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in Proc. 10thACM Workshop Artif. Intell. Secur. AISec, pp. 27 - 38, Aug, 2017

[5] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," arXiv preprint arXiv:1712.05526. Dec, 2017

[6] T. Gu, B. Dolan-Gavitt and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," arXiv preprint arXiv:1708.06733, Aug, 2017.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural netsworks," arXiv preprint arXiv:1312.6199. Feb, 2014.

[8] I. Goodfellow, J. Shlens and C. Szegedy, "Explaining and Harnessing Adversarial Examples," In Proc. International Conference on Learning Representations, May, 2015.

[9] A. Kurakin, I. Goodfellow and S. Bengio, "Adversarial examples in the physical world," In Artificial intelligence safety and security. Chapman and Hall/CRC. pp. 99-112, Aug, 2018

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, Jun, 2017

[11] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," In Proc. IEEE Symposium Security and Privacy, pp. 39-57, May, 2017.

[12] 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, Jul, 2016.

[13] Y. .Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9185-9193, Jun, 2018.

[14] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pp. 1322 - 1333, Oct, 2015

[15] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealingmachine learning models via prediction APIs," In 25th USENIX security symposium (USENIX Security 16), pp. 601-618, Aug. 2016.

[16] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition, Vol. 84, pp. 317-331, Dec, 2018.

[17] G. Ryu, H, Park and D.Choi, "Adversarial attacks by attaching noise markers on the face against deep face recognition," Journal of Information Security and Applications, Vol. 60, pp. 1-11, Aug, 2021

[18] H. Hirano, A. Minagi and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," BMC medical imaging, Vol. 21, No. 1, pp. 1-13. Jan, 2021

[19] K. Eykholt, I. Evtimov, E. Fernandes, B.Li, A. Rahmati, C. Xiao, 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.

[20] N. Akhtar and A. Mian, "Threat of ad

versarial attacks on deep learning in computer vision: A survey." IEEE Access, Vol. 6, pp. 14410-14430. Feb, 2018

[21] 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.

[22] W. Xu, D. Evans and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," arXiv preprint arXiv:1704.01155, Dec, 2017

[23] S. Freitas, S. Chen, Z. Wang, and D. Chau, "Unmask: Adversarial detection and defense through robust feature alignment," In 2020 IEEE International Conference on Big Data, pp. 1081-1088, Dec, 2020.

[24] R. Timofte and L. Van Gool, "Sparse representation based projections." In Proceedings of the 22nd British machine vision conference-BMVC, pp.61-61, Sep, 2011.

〈저자소개〉

박 소 희 (Sohee Park) 학생회원
2018년 2월: 공주대학교 응용수학과 학사
2020년 2월: 공주대학교 융합과학과 석사
2020년 6월~2021년 9월: 한국교육학술정보원 전문원
2022년 2월~현재: 숭실대학교 소프트웨어학과 박사과정
〈관심분야〉 인증, 금융보안, 인공지능 보안, 머신러닝

김 승 주 (Seungjoo Kim) 학생회원
2019년 2월~현재: 숭실대학교 소프트웨어학부 학사과정
〈관심분야〉 컴퓨터공학, 정보보호, 빅데이터

윤 하 연 (Hayeon Yoon) 학생회원
2019년 3월~현재: 숭실대학교 전자정보공학부 전자공학전공 학사과정
〈관심분야〉 정보보호

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〉 인증, 개인정보보호, 이상거래탐지, 의료정보보안, 머신러닝