https://doi.org/10.33778/kcsa.2022.22.3.025

데이터 증강 기반의 효율적인 포이즈닝 공격 방어 기법★

전 소 은*, 옥 지 원**, 김 민 정**, 홍 사 라**, 박 새 롬***, 이 일 구****

요 약

최근 이미지 인식 및 탐지 분야에 딥러닝 기반의 기술이 도입되면서 영상 처리 산업이 활성화되고 있다. 딥러닝 기술의 발전과 함께 적대적 공격에 대한 학습 모델 취약점이 계속해서 보고되고 있지만, 학습 시점에 악의적인 데이터를 주입하는 포이즈닝 공격의 대응 방안에 대한 연구가 미흡한 실정이다. 종래 포이즈닝 공격의 대응 방안은 매번 학습 데이터를 검사하여 별도의 탐지 및 제거 작업을 수행해야 한다는 한계가 있었다. 따라서, 본 논문에서는 포이즌 데이터에 대해 별도의 탐지 및 제거 과정 없이 학습 데이터와 추론 데이터에 약간의 변형을 가함으로써 공격 성공률을 저하시키는 기법을 제안한다. 선행연구에서 제안된 클린 라벨 포이즌 공격인 원샷킬 포이즌 공격을 공격 모델로 활용하였고, 공격자의 공격 전략에 따라 일반 공격자와 지능형 공격자로 나누어 공격 성능을 확인하였다. 실험 결과에 따르면 제안하는 방어 메커니즘을 적용하면 종래 방법 대비 최대 65%의 공격 성공률을 저하시킬 수 있었다.

Efficient Poisoning Attack Defense Techniques Based on Data Augmentation

So-Eun Jeon*, Ji-Won Ock**, Min-Jeong Kim**,
Sa-Ra Hong**, Sae-Rom Park***, Il-Gu Lee****

ABSTRACT

Recently, the image processing industry has been activated as deep learning-based technology is introduced in the image recognition and detection field. With the development of deep learning technology, learning model vulnerabilities for adversarial attacks continue to be reported. However, studies on countermeasures against poisoning attacks that inject malicious data during learning are insufficient. The conventional countermeasure against poisoning attacks has a limitation in that it is necessary to perform a separate detection and removal operation by examining the training data each time. Therefore, in this paper, we propose a technique for reducing the attack success rate by applying modifications to the training data and inference data without a separate detection and removal process for the poison data. The One-shot kill poison attack, a clean label poison attack proposed in previous studies, was used as an attack model. The attack performance was confirmed by dividing it into a general attacker and an intelligent attacker according to the attacker's attack strategy. According to the experimental results, when the proposed defense mechanism is applied, the attack success rate can be reduced by up to 65% compared to the conventional method.

Key words : Adversarial attack, Poisoning attack, One-shot Kill Poison Attack, Data Augmentation

접수일(2022년 08월 31일), 게재확정일(2022년 9월 15일)

★ 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2020R1F1A1061107)과 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원(P0008703, 2022년 산업혁신인재성장지원사업), 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음 (IITP-2022-RS-2022-00156310).

* 성신여자대학교 미래융합기술공학과 석사과정(주저자)
** 성신여자대학교 미래융합기술공학과 석사과정(공동저자)
*** 성신여자대학교 융합보안공학과/미래융합기술공학과 조교수(공동저자)
**** 성신여자대학교 융합보안공학과/미래융합기술공학과 조교수(교신저자)

1. 서 론

최근 인공지능 기술의 활용이 증가하면서 시장 규모 또한 커지고 있다. 특히 컴퓨팅 파워가 개선되고 머신러닝, 딥러닝 학습 알고리즘이 개발되면서 급격한 속도로 인공지능 기술이 발전하고 있다[1]. 빅마켓리서치에서 분석한 통계자료에 따르면, 2025년 머신러닝 시장이 3,980달러까지 성장할 것을 예측했다. 또한, 그랜드뷰리서치에 따르면, 딥러닝 시장 규모가 2030년까지 연평균 성장률이 34.3% 증가할 것으로 예측한 바 있다.

하지만, 딥러닝 기술이 널리 활용되면서 보안 위협도 함께 증가하고 있다. 특히 인공지능의 취약점을 이용해 의도적으로 잘못된 판단을 유발하는 적대적 공격이 빈번하게 발생하고 있다.

적대적 공격이란, 딥러닝 모델의 내부적 취약점을 이용하여 만든 특정 노이즈 값을 주입하여 학습 모델의 오분류를 유도하는 공격 방식이다. GAN(Generative Adversarial Network)을 개발한 이안 굿펠로우에 따르면, 사진에 매우 적은 잡음을 넣는 것만으로 최적의 신경망을 속일 수 있음을 보인 바 있다[2]. 또한, 2017년 7월 워싱턴대·미시건대 공동연구진은 '정지' 교통판에 스티커를 붙이자 자율주행 자동차가 '속도제한 시속 45마일' 표지판으로 오인식했다는 연구가 보고되었고[3], 중국의 킨시큐리티 연구소가 도로에 작은 점 3개를 칠하고 자율주행 자동차를 주행하도록 했을 때, 자율주행차가 차선을 잘못 인식하고 맞은편 차선으로 역주행하였던 연구 결과가 보고된 바 있다[4]. 하지만 종래의 적대적 공격은 주로 공격자가 모델의 학습 시점에 개입이 불가하다는 가정으로 테스트 시점에 공격을 수행하는 유형을 고려하였고, 학습 시점에 개입하는 포이즈닝 공격은 최근에 연구가 이루어지기 시작하여 연구가 미흡한 실정이다. 가령, 인터넷 공간에 공격자가 생성한 포이즌 데이터를 업로드 해두면, 사용자가 이를 크롤링하여 학습 모델에 쉽게 공격이 수행될 수 있어 위험성이 크다. 하지만, 종래의 연구에서는 포이즈닝 공격을 탐지하기 위해서 학습 데이터를 매번 검사하여 이상치를 제거하는 방법을 적용했다. 이는 학습 데이터에 대해 매번 검사 및 제거와 같은 별도의 작업이 필요하며 공격자가 최소한의 이상치를 가할 경우, 공격을 우회하여 수행할 수 있다는 한계점이 있다. 따라서 본 논문에서는 포이즌 데이터에 대해 별도의 탐지 및 제거 작업 없이 학습 데이터와 추론 데이터에 약간의 변형을 가함으로써 공격 성공률을 저하시키는 방법을 제안한다.

본 논문에서는 공격자가 별도로 학습용 데이터에 접근하지 못하더라도 학습 시점에 개입하여 공격을 수행할 수 있는 클린-라벨 공격(clean-label attack)인 원샷킬 포이즌 공격을 모델링하여 분석한다. 이 공격 방식은 포이즌 된 이미지를 공격자가 웹사이트에 업로드하기만 해도 사용자가 데이터를 크롤링하여 학습을 수행하면 공격이 쉽게 성공할 수 있어 공격 파급력이 크다는 문제점이 있다. 본 논문에서는 이 공격을 방어하기 위한 데이터 증강 기반의 효과적인 대응 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 적대적 공격 탐지 및 대응에 대한 선행연구를 분석하고, 3장에서는 원샷킬 포이즌 공격에 대해 설명한다. 4장에서는 데이터 증강 기법을 이용한 원샷킬 포이즌 공격에 대한 방어 기법에 대해 서술하며, 5장에서는 제안하는 방어 모델을 실험을 통해 검증하고, 6장에서 결론으로 마무리한다.

2. 관련 연구

최근 적대적 공격과 관련한 연구가 다양하게 이루어지고 있다. 본 장에서는 적대적 공격 탐지 및 대응 방법에 관한 선행 연구 동향을 분석한다.

Henry Chacon et al.[5]의 선행연구는 주어진 모델이 포이즌 피처가 포함되어 학습되었는지 탐지하는 알고리즘을 제안하고, CNN(Convolutional Neural Network) 모델을 활용하여 포이즈닝 공격이 모델의 매개 변수의 경계를 증가시키는 방법을 연구했다. 선행연구[5]는 정상 모델에서 임베딩 계층의 매개변수 분포의 경계와 적대적 피처로 학습된 모델에 대한 매개변수의 경험적 분포를 비교하는 방법을 제안하였다. Neehar Peri et al.[6]의

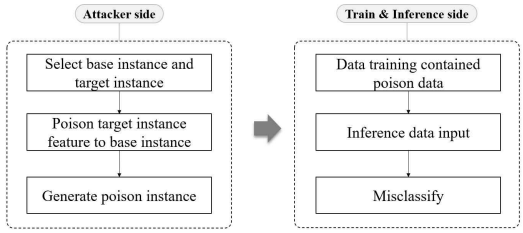
선행연구는 클린 라벨 포이즈닝 공격의 방어 기법인 Deep K-NN을 제안하였다. Deep K-NN은 피처 충돌[7]과 볼록 다포체 공격[8]을 고려한 방어 기법이며, 포이즌 데이터가 정상 데이터와 다른 피처 분포를 가지는 특성을 활용하여 학습 전에 특정 피처 공간에서 가장 가까운 이웃 k개를 위주로 탐지 및 제거하는 모델을 제안하였다. 하지만 선행연구 [5]와 [6]에서 제안한 기법은 학습 데이터에 대해 매번 별도의 탐지 및 제거 작업이 수행되어야 하고 공격자가 최소한의 적대적 피처로 포이즌할 경우 우회가 가능하다는 한계점이 있다.

Eun-na-rae Ko et al.[9]의 연구에서는 피처 선택, 피처 추출, 적대적 공격 탐지 단계로 구성되는 적대적 공격 탐지 및 분류 모델을 제안하였다. 선행연구[9]에서는 적대적 공격 대상 모델의 은닉층의 출력값에서 피처값을 추출하고, 이를 입력으로 하는 랜덤 포레스트 분류 모델을 구축하여 적대적 공격을 탐지 및 분류하는 방안을 제시하였다. 실험 결과 종래 연구 대비 3.02% 개선된 정확도를 보였다. 하지만 선행연구[9]는 적대적 공격의 대상 모델을 특정할 수 있을 때 사후 대응적인 측면에서 탐지 가능하다는 점에서 한계가 있다.

Weilin Xu et al.[10]은 피처 압축(feature squeezing)을 통해 적대적 예제를 탐지하는 연구를 수행했다. 피처 압축은 공격자의 공격 표면을 줄이기 위해 불필요한 입력 공간을 제거하여 적대적 예제를 생성할 가능성을 줄이는 방식이다. 선행연구[10]에서는 이미지 컬러 비트를 줄이는 방식과 공간 평활화의 두 가지 피처 압축 기법을 활용하였다. 이를 통해 원본 샘플의 예측 결과와 피처 압축이 적용된 모델의 예측 결과를 비교하여 적대적 공격을 탐지하는 방식이다. 선행연구[10]는 딥러닝 모델 자체를 수정하지 않고, 입력 이미지만을 변경하면서 적대적 예제를 탐지하는 모델을 제안하였다는 점에서 기여점이 있다. 하지만 공격자가 대상 모델에 대한 정보를 가지고 최소한의 노이즈로 적대적 예제를 생성할 경우 우회 가능성이 크다는 한계점이 존재한다.

3. 원샷킬 포이즌 공격 모델

본 연구에서는 인공지능 기반 학습 모델의 위협 모델로 선행연구[7]의 원샷킬 포이즌 공격을 활용한다. 원샷킬 포이즌 공격은 전이학습 환경에서 포이즌 데이터가 학습 데이터에 포함되었을 때 테스트 과정에서 포이즌 데이터가 입력되면 공격자가 의도한 분류 결과를 도출하는 공격이다. 이 공격은 인간의 눈으로 볼 때는 정상적인 이미지 데이터로 인식되기 때문에 정상 사용자가 라벨링을 하는 과정에서 오류를 유발하도록 한다. 즉, 공격자의 최소한으로 개입으로 큰 파급력을 초래할 수 있기 때문에 본 논문에서는 원샷킬 포이즌 공격을 공격 모델로 선정하였다. 그림 1은 이 공격 모델의 공격 수행 과정을 나타낸 것이다.



(그림 1) 원샷킬 포이즌 공격 과정

먼저 공격자는 포이즌 이미지 생성에 활용할 베이스 인스턴스와 타겟 인스턴스를 선정한다. 예를 들어, 고양이 이미지가 베이스 인스턴스일 때, 공격자는 다른 라벨의 타겟 인스턴스의 피처값을 베이스 인스턴스에 포함시켜서 포이즌 데이터를 생성한다. 이렇게 생성된 포이즌 데이터를 포함한 상태로 학습이 진행된다.

이후, 추론 단계에서 강아지 이미지의 타겟 인스턴스가 주어지면 타겟 인스턴스의 피처값이 포이즌 데이터인 고양이의 피처값과 같기 때문에 강아지 이미지를 고양이로 잘못 분류하게 된다. 즉, 인간의 눈으로 볼 때는 베이스 인스턴스와 같은 이미지로 보이지만 타겟 인스턴스와 동일한 피처값을 가지는 포이즌 데이터를 생성하고 테스트 단계에서 타겟 인스턴스가 입력값으로 들어오면 베이스 인스턴스로 오분류한다.

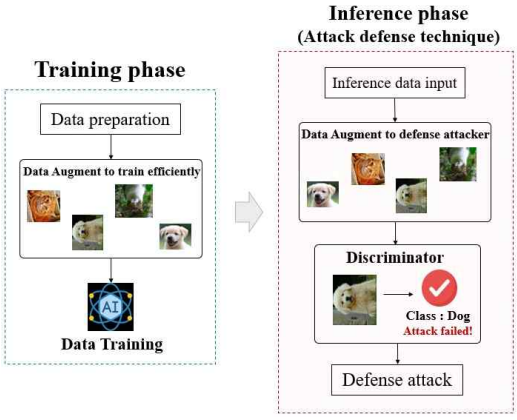
원샷킬 포이즌 공격 모델 이전의 적대적 공격은 학습 단계에 공격자의 개입이 불가하다는 가정으로 테스트 단계에서 공격을 수행했지만, 원샷킬 포이즌 공격 모델에서는 포이즌 데이터임에도 학

습 데이터에 레이블링이 적합하게 이루어진 것으로 보이는 상태에서 공격이 이루어질 수 있다는 점에서 공격의 파급력이 크다. 본 연구에서는 원샷킬 포이즌 공격을 방어하기 위한 모델을 제안한다.

4. 원샷킬 포이즌 공격에 대한 데이터 증강 기법 기반 방어 기법

본 장에서는 원샷킬 포이즌 공격의 방어 기법을 제안한다. 학습 시점에 개입하여 공격 데이터를 주입하는 원샷킬 포이즌 공격에 대응하기 위해 본 논문에서는 데이터 증강 (Data Augmentation) 기법을 활용한다. 데이터 증강 기법은 원본 데이터를 조작하여 변화를 가진 데이터를 생성하는 기법으로, 한정적인 데이터를 증강시킴으로써 성능을 높이고 오버피팅 문제를 해결하는데 활용한다[11].

그림 2는 제안하는 원샷킬 포이즌 공격의 방어 기법의 동작 구조를 나타낸 것이다.



(그림 2) 데이터 증강을 이용한 원샷킬 포이즌 공격의 방어 기법

효율적인 데이터 학습과 학습 시점에 개입하여 공격하는 공격자에 대응하기 위해 학습 데이터를 증강시켜 학습을 수행한다. 공격자는 데이터가 증강되어 학습되는 환경을 인지할 수 없으므로 원본 이미지 데이터로 포이즌 데이터를 생성하여 주입하게 된다. 따라서 공격자가 의도한 포이즌 수준으로 학습 모델이 중독되기 어렵다.

이후 추론 시점에서 공격자가 활용한 타겟 인스턴스가 입력되면, 추론 데이터도 증강시켜서 판별하는 알고리즘으로 동작한다. 이를 통해 학습 환경과 유사한 데이터의 구조로 판별할 수 있고, 모델의 분류 정확도를 향상시키는 동시에 공격자의 의도대로 오분류 할 가능성을 줄일 수 있다.

즉, 본 논문에서 제안하는 방어 기법은 공격자가 학습 데이터의 증강 유형을 인지하지 못하는 점을 이용하여 공격을 방어할 수 있다. 공격자가 학습 데이터의 증강 유형을 모르기 때문에 원본 이미지 데이터로 포이즌 이미지를 생성함으로써 공격자가 의도한 포이즌 라벨로 학습 수행이 불가하게 된다.

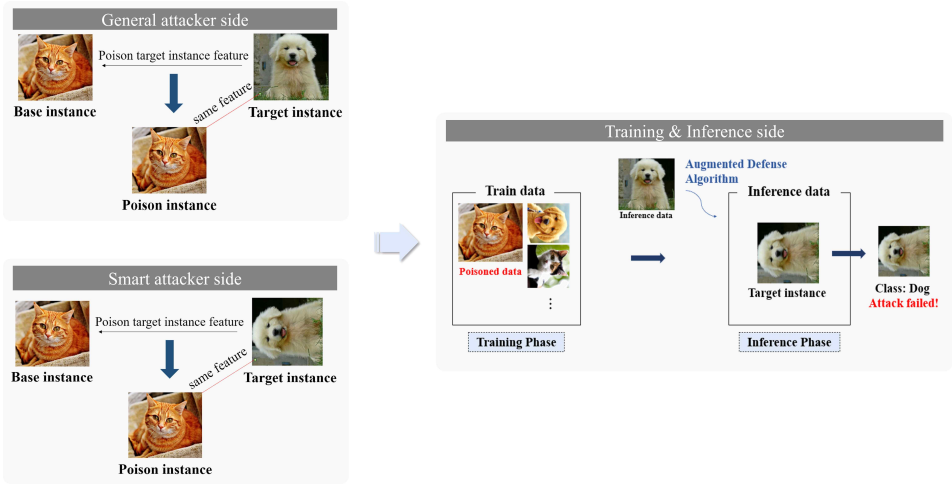
5. 성능 평가

5.1 실험 환경

원샷킬 포이즌 공격 실험 환경을 구축하기 위해서 깃허브에 공개된 코드[12]와, 학습 데이터가 비교적 적은 환경에서도 효율적으로 공격을 저하시킬 수 있음을 입증하고자 simple dog and cat dataset을 활용했다[13]. 이 데이터셋의 학습 데이터는 고양이, 강아지 라벨별로 각각 500개의 이미지 데이터로 구성되며, 테스트 데이터와 검증 데이터는 라벨별로 100개의 이미지 데이터로 구성된다. 학습 데이터가 부족할 때 효율적인 pre-trained ResNet 모델을 활용했으며, 10 에포크로 100번의 공격 시도가 발생하는 실험 환경에서 성능을 평가했다.

또한, 공격자의 공격 전략에 따른 공격 성공률을 확인하였다. 제안하는 방어 기법이 동작하는 것을 인지하지 못하고 공격을 수행하는 일반 공격자와 제안하는 방어 기법이 동작하는 것을 인지하고 공격을 우회하는 지능형 공격자 환경에서 제안하는 방어 기법의 성능을 평가하였다.

그림 3은 일반 공격자(General attacker)와 지능형



(그림 3) 공격 전략에 따른 공격 과정

공격자(Smart attacker)의 공격 과정을 나타낸 것이다. 일반 공격자는 방어 기법의 동작을 인지하지 못하기 때문에 타겟 인스턴스를 증강시키지 않은 채로 포이즈닝하고, 지능형 공격자는 학습 모델의 데이터 증강 환경을 인지하고 우회하기 위해 타겟 인스턴스도 학습 환경의 증강 유형에 맞게 변형시킨다. 이때 지능형 공격자는 모델에 여러 번의 사전 공격을 시도하여 모델의 학습 환경에 적용된 증강 유형을 인지한 공격자인 것으로 가정한다.

공격자는 베이스 인스턴스에 타겟 인스턴스를 5,000번을 반복하여 포이즈닝하는 환경이다. 피처 공간은 포이즌 인스턴스의 피처가 타겟 인스턴스의 피처와의 피처 거리를 최소화하고, 입력 도메인은 포이즌 인스턴스와 베이스 인스턴스가 같은 픽셀을 가지도록 MSE(Mean Squared Error) 평가지표를 이용해서 업데이트한다. 본 실험에서 데이터 증강 방식으로 회전 방식을 선정하였고, 회전 각도는 공격 라운드마다 1°~360° 범위에서 랜덤으로 회전되도록 하였다.

표 1은 모델별 증강 수행 여부를 정리한 표이다. Conventional scheme은 공격에 대한 방어 메커니즘이 적용되지 않은 일반적인 원샷킬 포이즌 공격 유형이다. 따라서 모든 데이터가 증강되지 않고 인지할 방어 알고리즘이 적용되지 않은 모델이므로 일반 공격자만 개입하는 환경이다. 제안하는 모델은 학습 데이터만 증강하거나, 학습 데이터와 추론 데이터를 함께 증강시키는 것에 따라 Proposed scheme_A, Proposed scheme_B로 나누어 성능을 비교하였다. Proposed scheme_A는 추론 데이터에는 증강 기법이 적용되지 않은 모델로 학습 환경의 효율성과 공격이 개입되는 시점만 고려한 것이고, Proposed scheme_B는 학습 데이터를 포함한 추론 시점에 입력되는 데이터도 증강하여 공격 성공률을 저하시키는 모델이다.

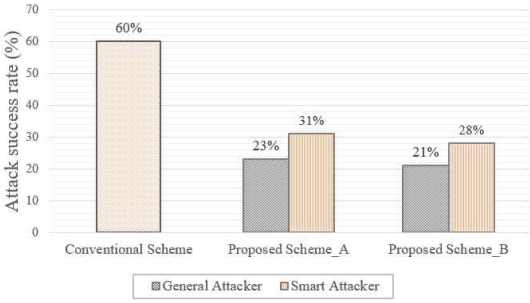
<표 1> 모델별 실험 환경

모델	공격자 유형	Augment 대상		
		학습 데이터	공격 데이터	추론 데이터
Conventional Scheme	General attacker	X	X	X
Proposed Scheme_A	General attacker	O	X	X
	Smart attacker	O	O	X
Proposed Scheme_B	General attacker	O	X	O
	Smart attacker	O	O	O

자원이 제약된 경량 장치는 증강 알고리즘을 최소화한 Proposed scheme_A를 활용하여 공격에 대응할 수 있고, 비교적 자원이 충분한 장치는 공격을 최적으로 방어할 수 있는 Proposed scheme_B를 활용하여 대응할 수 있다.

5.2 실험 결과 및 분석

그림 4는 원샷킬 포이즌 공격을 재구현하여 공격을 수행했을 때의 종래 공격 모델의 성능과 제안하는 방어 기법을 적용했을 때 공격 성공률을 비교한 결과이다.



(그림 4) 원샷킬 포이즌 공격 방어 기법의 성능 비교 결과

본 논문에서 제안하는 방어 모델이 적용되지 않은 종래 원샷킬 포이즌 공격 환경에서는 60%의 공격 성공률을 보였다. 이는 데이터가 증강되지 않은 환경에서 공격자가 학습 환경의 데이터와 같이 원본 이미지로 포이즌 데이터를 생성했기 때문에 오분류를 유발하는 비율이 높은 것을 알 수 있다. 이에 반해 방어 모델이 적용된 Proposed scheme에서는 공격 성공률을 50% 이상 저하시킬 수 있었다. 특히 학습 데이터와 추론 데이터에 증강 메커니즘을 모두 적용한 Proposed scheme_B에서 공격 성공률이 각각 21%, 28%로 더 효과적임을 확인할 수 있었다. 또한, 학습 환경의 효율성과 공격이 개입되는 시점만 고려한 모델인 Proposed scheme_A도 23%, 31%로 공격 성공률이 저하됨을 확인할 수 있었다. 특히, 학습 환경의 증강 메커니즘을 인지하지 못하는 일반 공격자는 종래 공격 모델 대비 각각 61.7%, 65%만큼 저하되었고, 학습 데이터의 증강 유형을 인지한 지능형 공격자는 그보다 증가된 공격 성공률을 보였다. 그럼에도 종래 공격 모델 대비 각각 48.3%, 53.3%의 공격 성공률이 저하됨을 확인하였다.

이는 공격자가 학습 시점에 개입하더라도 내부의 학습 상황은 알 수 없다는 점을 활용하였다. 정상 학습 데이터를 증강시켜서 학습하고, 추론 시점에 입력되는 데이터도 증강함으로써 공격자가 의도한대로 오분류를 유발하는 것이 어렵기 때문에 공격 성공률이 크게 저하된다. 두 모델 모두 간단한 증강 알고리즘을 통해 효율적으로 공격 성공률을 저하시켰으며, 복잡도가 최소화 된 증강 알고리즘이 적용된 Proposed Scheme_A에서도 공격 성공률을 크게 저하시킬 수 있음을 입증하였다. 또한, 학습 데이터만 증강시킨 환경에서도 공격 이미지를 제외한 정상 이미지는 모두 다양한 각도로 회전된 이미지가 학습되기 때문에 정방향 이미지가 추론 시점에 입력되더라도 공격 성공률을 종래 모델보다 저하시킬 수 있다. 또한, 일반 공격자는 학습 데이터와 추론 데이터의 증강 환경을 인지하지 못한다는 점에서 정상 데이터와 달리 원본 이미지로 학습되기 때문에 공격 성공률이 크게 저하된다. 이에 반해, 지능형 공격자는 학습 데이터의 증강 유형을 인지하고 공격 이미지도 회전하여 포이즌하지만, 다양한 각도로 회전하여 학습되는 환경으로 인해 공격 성공률이 종래 모델 대비 저하되는 결과를 보였다.

6. 결 론

인간이 구분하기에 어려운 정도로 데이터에 미세한 변화를 부여하는 것만으로도 모델의 판단 결과의 변화가 발생할 수 있다. 이로 인해 발생하는 오인식을 적대적 공격이라고 하며, 원본 이미지에 왜곡을 추가하여 모델의 오인식을 유도하는 방법이 있으므로 적대적 공격이 유리한 측면이 있다. 예측하기 어려운 딥러닝 결과의 오분류를 유도함으로써 잠재적인 위험성을 가지며, 의료기기나 자율주행자동차 분야 등에서 적대적 공격이 발생한다면 심각한 문제가 될 수 있어 대응책이 필요하다. 이를 방어하기 위해 원본 이미지의 정확도를 유지하면서 적대적 이미지를 잘 인식하거나 탐지하는 것이 중요하다.

따라서 본 연구에서는 적대적 공격 모델 중 학습 모델을 공격하는 포이즈닝 공격을 방어하기 위해 데이터 증강을 이용한 공격 방어 알고리즘을 제안하였

다. 추론 시점에 입력되는 추론 데이터를 무작위로 회전시킴으로써 원샷킬 포이즌 공격에 효율적으로 대응할 수 있음을 입증하였다. 알고리즘의 성능을 입증하기 위해 공격자의 공격 전략에 따라 일반 공격자와 지능형 공격자로 나누어 공격 성능을 실험하였다. 이 방어 모델을 통해서 일반 공격자 환경과 지능형 공격자 환경에서 종래 공격 모델[7] 대비 원샷킬 포이즌 공격 성공률이 각각 최대 65%, 53% 저하됨을 입증하였다.

향후 연구에서는 원샷킬 포이즌 공격에 선제적으로 대응하기 위해 학습 단계부터 데이터 증강을 적용하는 방어 기법들을 연구하고, 회전 이외에도 다양한 데이터 증강 기법을 추론 시에 적용하여 예측 결과의 앙상블 기법을 연구할 계획이다.

참고문헌

[1] D. Lee, "Security threat verification and defense techniques in artificial intelligence for image processing(Masters thesis)", Chungnam National University, 2021.

[2] I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and Harnessing Adversarial Examples", International Conference on Learning Representations(ICLR), 2015..

[3] K. Eykholt,, I. Evtimov, E. Fernandes, "Robust Physical-World Attacks on Deep Learning Visual Classification.", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1625-1634, 2018.

[4] H. C. Kwon, S. J. Lee, J. Y. Choi, B. H. Chung, S. W. Lee, J. C. Nah, "Security Trends for Autonomous Driving Vehicle", Electronics and Telecommunications Trends, Vol. 33, No. 1, pp. 78-88, 2018.

[5] H. Chacon, S. Silva and P. Rad, "Deep Learning Poison Data Attack Detection", IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 971-978, 2019. doi: 10.1109/ICTAI.2019.00137.

[6] P. Neehar, G. Neal and Huang, W. Ronny, F. Liam, Z., Chen, F. Soheil, G. Tom, D. John "Deep k-NN Defense Against Clean-Label Data Poisoning Attacks," Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23 - 28, 2020, Proceedings, Part I, vol 12535. 2020. https://doi.org/10.1007/978-3-030-66415-2_4.

[7] A. Shafahi, W. R Huang., M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks", Advances in neural information processing systems, Vol. 31, 2018.

[8] C. Zhu, W.R. Huang, H. Li, G. Taylor, C. Studer, T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," International Conference on Machine Learning. pp. 7614 - 7623, 2019.

[9] E. N. R. Ko, and J. S. Moon, "Adversarial Example Detection and Classification Model Based on the Class Predicted by Deep Learning Model.", Journal of the Korea Institute of Information Security & Cryptology, Vol. 31, No. 6, pp. 1227-1236, 2021.

[10] W. Xu, D. Evans, Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," Symposium on Network and Distributed Systems Security, 2018.

[11] C. Shorten, T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning",. J Big Data Vol. 6, No. 60. 2019. https://doi.org/10.1186/s40537-019-0197-0.

[12] Dong bin Na, 'Poison-Frogs-OneShotKillAttack-PyTorch', https://github.com/ndb796/Poison-Frogs-OneShotKillAttack-PyTorch, 검색일: 2022. 8. 18.

[13] Dong bin Na, 'simple_dog_and_cat_dataset', https://github.com/ndb796/Poison-Frogs-OneShotKillAttack-PyTorch/tree/main/simple_dog_and_cat_dataset, 검색일: 2022. 8. 18.

〔저자소개〕

전 소 은 (So-Eun Jeon)
2021년 8월: 성신여자대학교 융합보안공학과 학사
2021년 9월~현재: 성신여자대학교 미래융합기술공학과 석사과정
email : 220214016@sungshin.ac.kr

옥 지 원 (Ji-Won Ock)
2022년 8월: 성신여자대학교 융합보안공학과 학사
2022년 9월~현재: 성신여자대학교 미래융합기술공학과 석사과정
email : 220224011@sungshin.ac.kr

김 민 정 (Min-Jeong Kim)
2022년 2월: 성신여자대학교 융합보안공학과 학사
2022년 3월~현재: 성신여자대학교 미래융합기술공학과 석사과정
email : 220226033@sungshin.ac.kr

홍 사 라 (Sa-Ra Hong)
2021년 2월: 성신여자대학교 융합보안학과 학사
2021년 9월~현재: 성신여자대학교 미래융합기술공학과 석사과정
email : 220216140@sungshin.ac.kr

박 새 롬 (Sae-Rom Park)
2013년 2월 서울대학교 산업공학과 학사
2018년 2월 서울대학교 산업공학과 박사
2019년 9월~현재: 성신여자대학교 융합보안공학과/미래융합기술공학과 조교수
email : psr6275@sungshin.ac.kr

이 일 구 (Il-Gu Lee)
2003년 2월 서강대학교 전자공학과 학사
2005년 2월 KAIST 정보통신대학원 석사
2012년 2월 KAIST 지식재산대학원 석사
2016년 2월 KAIST 전산학부 박사
2017년 2월~현재: 성신여자대학교 융합보안공학과/미래융합기술공학과 조교수
email : iglee@sungshin.ac.kr