AI 시대의 보안 및 프라이버시 이슈

최대선 숭실대학교 소프트웨어학부 교수

1. 머리말

영상, 음성, 언어 데이터 등을 딥러닝 기술을 바탕으로 처리하는 인공지능(AI)이 여러 분야에서 활용되고 있다. 활용 효과는 적용 분야에 따라 논란이 있지만 앞으로 더욱 많은 분야에서 더욱 많이 활용되라는 점에는 이견이 없는 듯하다. 이에 따라 AI의 보안과 신뢰성, 역기능에 대한 관심이 고조되고 있다.

AI의 보안은 AI기술을 활용해 악성코드탐지, 침입탐지 같은 기존의 보안 문제를 해결하는 'AI를 이용한 보안(AI for security)'과 시스템 보안이나 소프트웨어 보안과 별개로 AI 모델이 갖는 별도의 취약점을 방어하기 위한 'AI를 위한 보안(security for AI)'으로 나눌 수 있다. AI의 신뢰성 분야는 특정 데이터 집합으로 학습된 AI가 도메인이나 시점 등이 다른 데이터 환경에서 성능을 유지하는 문제와, AI의 판단 이유를 분석하여 AI 오동작 등의 원인을 파악하기 위한 설명가능한 AI(XAI, eXplainable AI)를 포함한다. AI의 역기능에는 딥페이크와 같이 AI를 이용해 사람을 속이는 문제와, 학습데이터 편향에 의한 인종차별과 같은 AI 오동작 등이 포함된다. 한편, 자율주행 자동차의 오동작 같은 문제는 AI 안전성(Safety) 분야에서 따로 다뤄지고 있다.

프라이버시 이슈는 AI의 역기능 중에 하나지만, 따로 더 세분화할 필요가 있는 중요한 이슈다. 우선 AI 학습에 사용되는 데이터는 많은 경우 개인의 민감정보를 담고 있다. 프라이버시 보호를 위해 비식별 처리를 하면 AI 학습을 위한 데이터로서의 유용성이 크게 감소한다. 또 이미지와 영상, 생체 신호 등의 비정형 데이터는 개인이 식별되는 유형이 매우 다양하여 비식별화하는 방법이 확립되어 있지 않은 실정이다. 한편, 학습이 완료된 AI에서 학습데이터를 추출하거나, 특정인이 학습데이터에 포함되어 있는지를 판단하는 것이 가능하여 또 다른 프라이버시 문제가 되고 있다. AI 스피커 등 AI 기기를 통해 개인정보가 수집, 유출되는 경우도 있어 다양한 고려가 필요하다.

본고에서는 AI를 위한 보안과 AI 학습데이터 프라이버시 이슈를 중점적으로 살펴본다.

2. AI를 위한 보안

2.1 AI에 대한 적대적 공격

AI 모델이 구동되는 OS 등 시스템 보안과 소프트웨어 보안과 별개로 AI 모델은 고유의 보안 취약점을 갖고 있다. 2014년 AI 모델에 대한 기만공격이 알려진 이래 공격자 측면에서 여러

가지 공격이 가능하다는 것이 연구되어 왔다. AI 모델에 대한 공격은 AI의 학습단계와 활용단계에 따라 나눌 수 있는데, 학습단계에서는 학습데이터에 오염데이터를 주입하여 모델의 정확도를 떨어뜨리는 오염 공격과 특정 패턴을 포함한 이미지를 특정 클래스로 분류하는 백도어 공격이 가능하다. 학습된 AI 모델을 활용하는 단계에서는 데이터를 변조하여 모델의 오분류를 유도하는 기만 공격, 학습에 사용된 데이터 복원이나 멤버십 추론, 모델 복제 등의 공격이 가능하다.

오염 공격(Poisoning attack)은 학습데이터 중 개 사진에 고양이 레이블을 붙이는 것 같은 오염데이터를 포함하여 학습된 모델의 정확도를 떨어뜨리는 공격이다. 공격자의 목표는 최소한의 오염데이터 비율로 최대한 정확도 저하를 유도하는 것이다. 오염 공격이 발생하는 시나리오는 세 가지 정도가 알려져 있다. 첫째, 아웃소싱을 통해 학습데이터를 구축하거나 AI 모델을 개발할 때 오염데이터를 주입하는 경우이다. 둘째, 이미 학습된 모델에 자신의 학습데이터를 추가하여 일종의 커스터마이징을 수행하는 전이학습(transfer learning)에서 사용되는 스승 모델을 오염데이터로 학습시켜 배포할 수 있다. 셋째, 분산된 클라이언트가 로컬모델을 학습한 후 가중치(weights) 집합 등 학습 파라미터를 중앙 서버로 전송하여 글로벌 모델을 만드는 연합학습에서, 특정 클라이언트가 오염데이터를 학습하여 생성된 파라미터를 서버로 전송하는 경우이다.

백도어 공격(backdoor attack)은 학습데이터에 트리거(trigger)라고 불리는 특정한 패턴을 포함하여 학습하고, 활용단계에서 트리거를 포함한 입력 데이터에 대해 특정 클래스로 분류하도록 하는 공격을 의미한다. 다른 방식으로, 학습데이터에 트리거를 포함하지 않고 모델의 특정 파라미터를 수정하여 트리거가 포함된 입력 데이터에 반응하게 할 수도 있다. 이러한 백도어 공격은 오염 공격과 마찬가지로 아웃소싱 개발이나 전이학습 때 발생할 수 있다. 트리거의 종류는 특정 위치에 특정 패턴을 고정적으로 포함하는 정적(static) 트리거와 객체의 경계선 역영에 특정 색상을 주입하는 식으로 이미지에 따라 트리거가 달라지는 동적(dynamic) 트리거가 있다. 이미지의 밝기를 영역에 따라 달리하거나, 다른 이미지 또는 텍스트와 합성하거나, 이미지를 회전시키거나 형태를 약간 찌그러뜨리는 종류의 트리거도 있다. 사람 얼굴의 경우 특정 악세사리를 착용한 사진이나, 딥페이크 등에 활용되는 이미지 인페이팅(image inpainting) 기술을 활용한 이미지 합성 기법이 사용되기도 한다. [그림 1]은 여러 종류의 트리거의 예시를 보여준다.

출처: 자체

[그림 1] 여러 가지 백도어 트리거 예시

기만 공격(evasion attack)은 대표적인 활용단계 공격이다. 학습을 마친 AI 모델의 활용단계에 입력되는 데이터를 변조하여 오분류를 유도하는 공격인데, 이때 변조된 데이터를 적대적 예제

(adversarial example)라고 부른다. 공격을 위해 변조되는 폭이 크지 않아서 사람이 인식하기에는 차이가 없어 보이지만 AI 모델은 전혀 다른 클래스로 분류하게 된다. 이미지, 음성, 텍스트 등 거의 모든 종류의 데이터에 대한 적대적 공격 연구가 이루어졌다. AI 스피커에게 음악을 들려주었는데 음성 명령으로 인식하여 수행하거나, 자율주행차가 도로 표지판을 오인식하게 하여 사고를 유발하는 등의 시나리오가 가능하다. 디지털 데이터에 대한 변조뿐 아니라, 실제 도로 표지판이나 사람 얼굴에 스티커를 부착하는 물리적 공격(physical attack)도 가능하다. 또한 악성코드 탐지, 침입 탐지 등 보안 기능을 위한 AI 모델을 속일 수 있도록 실행파일이나 네트워크 패킷을 변조하는 공격도 연구되었다.

기만 공격에서 공격자가 AI 모델에 대한 모든 정보를 알고 있고 AI의 출력 결과에 따라 입력 데이터를 변조할 수 있는 화이트 박스 환경에서는 100%에 가까운 공격 성공률(의도한 다른 클래스로의 오분류)을 보인다. 공격 대상 모델에 자유로운 접근이 불가능한 블랙박스 상황에서도 유사한 기능을 수행하는 AI 모델을 생성하여 이를 대상으로 모의 공격을 통해 유효한 변조 데이터를 획득한 뒤, 이를 이용해 실제 모델을 공격하는 전이 공격(transfer attack)도 가능하다. 다만 공격 성공률은 화이트박스의 경우보다는 떨어진다.

하나의 AI 모델을 개발하기 위해서는 학습데이터 구축과 학습과정에서 많은 시간과 비용이 필요하다. 그런데 이렇게 개발된 AI 모델을 쉽게 복제할 수 있는 공격이 연구되었다. 딥러닝 모델의 경우 네트워크 구성과 가중치 집합 정보만 있으면 AI 모델을 복제할 수 있다. 하지만 이러한 내부 정보에 대한 접근 없이 외부에서 정상적으로 AI 모델을 이용하는 과정을 통해 모델 복제가 가능한 것은 큰 위협이 아닐 수 없다. 연구된 공격은 여러 개의 데이터를 분류하도록 AI 모델에 질의하고 AI의 응답 결과를 분석하여 복제 모델을 만드는 방식이다.

유사한 방식으로 활용단계의 AI 모델에 질의 후 응답 결과를 분석하여 AI 모델 학습에 특정 데이터가 포함되었는지 판단하는 멤버십 추론(membership inference) 공격과 학습에 사용된 데이터를 복원하는 도치(inversion) 공격도 가능하다. 멤버십 추론과 도치 공격은 프라이버시 침해를 야기할 수 있다.

2.2 AI 모델 보안

AI 모델 보안은 앞 절에서 언급한 다양한 적대적 공격들에 대한 방어를 의미한다. 다른 보안 이슈와 마찬가지로 AI 모델 보안 방안도 기술적 방안과 절차적 방안으로 나눌 수 있다. 기술적 방안은 학습데이터 오염 공격에 대해서는 학습데이터 자체를 정제하거나, 오염된 데이터를 학습하여도 오분류를 하지 않는 강건한 모델이 되도록 반복적으로 재훈련하는 것이다. 절차적 대응은 아웃소싱이나 전이학습 등에 대한 보안 및 신뢰관리를 통해 보안을 확보하는 것이다.

백도어 공격에 대한 방어에는 학습된 모델의 백도어 포함 여부에 대한 분석과 입력된 데이터의 트리거 포함 여부 탐지, 백도어의 영향을 무력화하기 위한 데이터 변환, 백도어가 포함되어도 영향을 미치지 않도록 하기 위한 모델 재학습 등의 방법이 있다. 절차적 대응 방안은 같은 학습 단계 공격인 학습데이터 오염공격의 경우와 같다.

기만 공격은 공격 방법에 대해 많이 연구된 것처럼 방어 기법에 대해서도 많은 연구가 있었는데 크게 적대적 학습, 필터링, 탐지의 세 가지로 분류할 수 있다. 첫째, 적대적 학습

(adversarial training)은 적대적 예제를 생성하여 이를 학습데이터에 포함하고 적대적 예제도 올바르게 분류할 수 있도록 한다. 새로운 적대적 예제 또는 방어가 적용된 모델을 대상으로 하는 적응 재공격(adaptive attack)에 대응하기 위해 모델 내의특징(feature) 인식 단계에서 원본 이미지와 적대적 예제의 특징이 유사하게 인식되도록 학습하기도 하는데, 이는 전이 공격 성공률을 높이려는 공격 기법에서 아이디어를 얻어 방어에 사용하는 방안이다. 둘째, 필터링은 변조가 적용된 데이터를 재변조해서 공격 효과를 약화시키는 방법이다. 기만 공격을 위한 입력 데이터 변조는 사람에게 인식되지 않기 위해 최소화된 변조가 이뤄지기 때문에, 단순하게는 색상별로 8비트를 사용하는 이미지 데이터에서 최하위 비트를 삭제하는 것만으로도 변조의 영향을 없앨 수 있다.

셋째, 탐지는 적대적 예제가 정상 데이터와 다른 특징을 찾아내어 적대적 예제임을 인식하는 것이다. 예를 들면, 적대적 예제는 최소한의 변조를 위해 분류 결과에서 공격 목표인 타 클래스 인식확률이 원래 클래스보다 높아지면 변조를 중단하게 된다. 이 결과 보통 특정 클래스에 편중되었던 클래스별 확률이 원본 클래스와 타 클래스에 분산되어 나타나게 된다. 이같은 특징을 바탕으로 적대적 예제를 인식할 수 있다.

AI 모델 보안도 다른 보안 문제와 마찬가지로 창과 방패의 싸움이다. 새로운 공격기법이 발표되면, 얼마 후 이에 대한 방어기법이 발표되지만, 방어 기법이 적용된 모델을 하나의 타겟으로 놓고 다시 공격(적응 재공격)해보면 방어 성능이 매우 떨어지는 결과가 나온다. 얼마 후에는 방어기법을 거의 완전히 무력화시킬 수 있는 새로운 공격 기법이 등장하게 된다. 그래서 현재까지는 공격 쪽이 다소 유리한 입장이다. 미국 국방성 R&D 관리기구인 DARPA에서는 AI 모델 보안을 위한 GARD(Guaranteeing AI Robustness against Deception) 프로그램을 운영하는 중이다. 여기서는 주로 기만 공격을 방어하기 위한 연구를 지원하고 있는데, 이 연구의 목표는 기만 공격에 대한 완전한 방어가 아니라 새로운 공격이나 적응 재공격에 대해 완전히 무력화되지 않는 방어 기술을 개발하는 것이다. 이렇게 AI 모델 보안 이슈는 점진적 개선을 위한 지속적 연구가 필요한 분야이다.

3. AI 프라이버시 이슈

3.1 학습데이터의 프라이버시

AI 학습에 쓰이는 많은 데이터는 개인들이 금융이나 의료 등 어떤 서비스를 이용한 결과 생성된 것이다. 이러한 데이터는 개인정보이므로 AI 학습에 활용하기 위해서는 AI의 목적과 기능을 명확히 하고 정보 주체의 동의를 받아야 한다. 하지만 많은 사용자에게 동의를 받는 것은 용이하지 않다. 개인정보보호법 등 데이터3법이 개정되면서 가명처리된 데이터는 제한된 목적에 한해 정보 주체의 동의를 받지 않고 사용할 수 있는 법적 근거가 생겼다. 또한 재식별이 불가능하게 만든 익명처리된 데이터는 더 이상 개인정보가 아니므로 제한 없이 활용할 수 있게 되었다.

법 개정 이후 금융권을 중심으로 익명 및 가명처리된 데이터 활용이 활발히 진행되고 있다. 현재 활용되는 데이터는 데이터베이스에 테이블 형태로 저장된 정형 데이터로서 이에 대한 가명처리 및 익명처리 기법은 이론적으로 확립되어 있다. 데이터의 익명처리는 주민번호, 전

화번호, 이름 등 식별자를 제거하는 단계, 그 자체로는 개인식별이 안되지만 몇 개의 필드를 조합(예, 특정지역의 특정 직업을 가진 35세의 남자)하면 개인을 식별할 수 있는 필드들을 처리(K-익명성적용 등)하는 단계 등으로 구성된다. 많은 실제 데이터들이 가명 및 익명 처리되면서 처리 노하우나 처리 수준에 대한 컨센서스 또한 확립되어 가고 있는 상황이다.

그런데 AI 학습에 많이 사용되는 영상, 이미지, 텍스트, 음성 등 비정형 데이터에 대해서는 처리 방법이 확립되지 않았다. 현재 영상, 이미지에 포함된 이름, 주민번호, 전화번호, 자동차 번호, 얼굴 등의 식별정보를 탐지하여 마스킹하는 정도의 수준이다. 데이터에 포함된 정보를 조합하거나 다른 정보와 결합하여 식별할 수 있는 문제를 다루는 단계에 이르지 못하고 있으며, 개별적 기술들이 단편적으로 연구 및 개발되고 있다.

자유로운 텍스트에서 직업명, 지역명, 사람 이름 등을 찾아내는 기술을 개체명 인식이라고 부른다. 한국어에서 개체명 인식은 형태소 분석을 통해 형태소 후보군이 인식된 다음, 주변 문맥을 보고 해당 형태소가 고유명사 즉 개체명 후보가 되는지 판단하는 방식으로 이뤄진다. 개체명 후보에 대해서 이것이 지명인지, 회사명인지, 사람 이름인지 문맥에 따라 판단하게 된다. TTA 표준에서는 146개의 세부 개체명 분류가 있고, 한국어 개체명 인식의 정확도는 대략 80% 수준이다. 인식된 개체명은 단순하게는 마스킹 처리를 할 수 있으나, 모든 개체명을 마스킹 처리하면 의미 있는 정보가 거의 남지 않게 된다. 인식된 개체명으로 익명처리를 제대로 하기 위해서는 해당 개체명이 누구의 정보인지 주체를 파악하고, 해당 정보 주체의 어떤 정보인지 파악해야(예, 지명의 경우 거주지인지 직장주소인지) 한다. 그래야 정형데이터처럼 이들의 조합을 통한 식별가능성을 판단할 수 있기 때문이다. 하지만 정보주체를 파악하는 일부터 쉬운 문제는 아니다.

어떤 이미지나 동영상이 있을 때 프라이버시 침해 요소는 전화번호 같은 고유식별 번호나 사람의 얼굴 등이 될 수 있다. 이미지나 동영상에서 고유식별번호를 탐지하는 것은 높은 정확도에 도달해 있고, 사람 얼굴 영역도 MTCNN(Multi Task CNN) 등 딥러닝 기술을 사용하여 탐지할 수 있다. 탐지된 얼굴 영역은 마스킹, 흐리게 만들기(bluring), 다른 사람 얼굴로 합성하기(inpainting) 등을 통해 익명처리할 수 있다. 하지만 얼굴을 탐지하고 처리한다고 해서 완벽한 익명처리가 될 수는 없다. 장소, 시간, 옷차림 등으로 얼굴이 드러나지 않은 유명인을 식별할 수 있었던 사례처럼 식별 가능성은 여전히 남아있기 때문이 다.

의료 분야에서는 X-ray, CT, MRI 등 의료 영상의 가명처리가 이슈이다. 국내 의료 분야에서는 어떤 이미지가 있을 때 이를 완벽히 익명처리하는 것이 불가능하다고 전제하고 있다. 따라서 의료 영상 처리는 가명처리라고 부른다. 가명정보는 다른 정보와 결합하여 식별이 가능한 것이고 익명정보는 다른 정보와 결합하여도 식별이 불가능하다는 법적인 정의에 비추어 타당한 전제로 볼 수 있다. 의료 영상의 가명처리는 얼굴부분의 영상에 대해 윤곽선을 찾아 제거하는 것을 주된 처리 방법으로 하는데, 윤곽선을 통해 얼굴을 복원하는 기술이 있기 때문에 이러한 처리가 필요한 것이다.

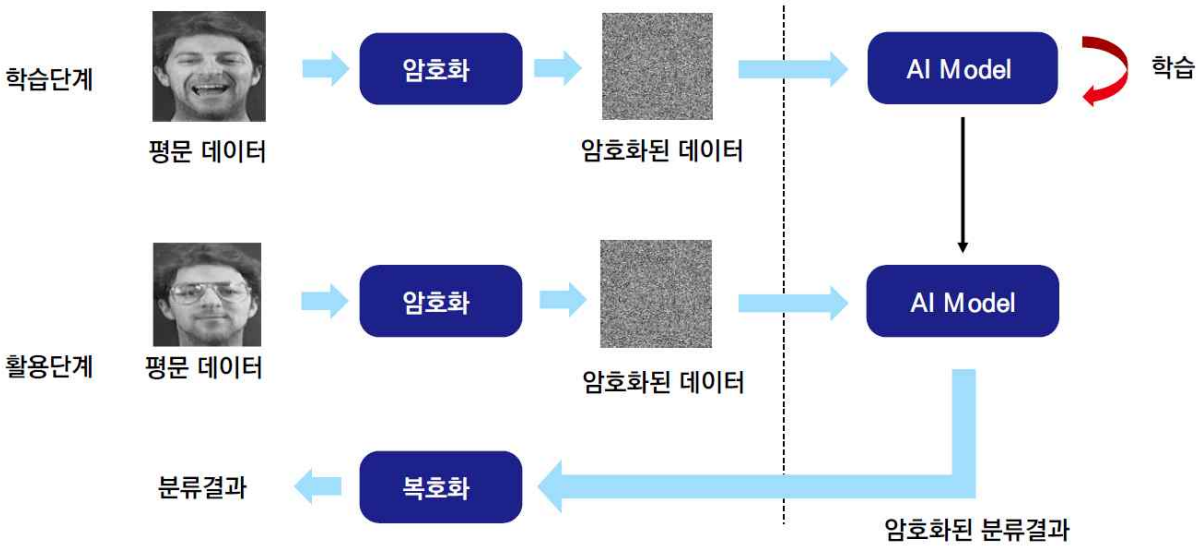
음성 데이터는 데이터 전체가 개인식별 가능성이 있다. 어느 영역을 마스킹하는 것은 효과가 없으며, 주파수 대역을 변경하는 등의 조치를 취해도 거꾸로 복원이 가능하다. 음성을 텍스트로 인식(speech to text)한 후 다시 텍스트를 음성으로 변환(text to speech)하는 방법을 생각

할 수 있는데, 이 경우는 본래 음성의 모든 특징이 사라지게 되므로 학습데이터로서의 의미가 있는지 판단해야 한다.

3.2 프라이버시 보존형 머신러닝

가명 및 익명처리를 통해 데이터의 유용성이 감소하나, 비정형 데이터에 대한 적절한 처리 방법은 아직 확립하지 못하였다. 이에 따라, 동형암호나 재현데이터 등의 대안이 모색되고 있다. 학계에서는 프라이버시 보존형 머신러닝(privacy preserving machine learning)이라는 분야로 관련 연구가 진행되어 왔다.

동형 암호(homomorphic encryption)는 암호화된 상태에서 수치연산 등이 가능한 기술이다. 학습데이터를 통해 파라미터를 계산하는 것이 AI 학습인데, 학습 데이터를 암호화하여 전송하고 암호화된 상태에서 학습하게 하는 것이다. 활용단계에서도 암호화된 데이터를 전송하여 처리 결과도 암호화된 상태에서 전송받으면 이를 복호화하여 사용한다. [그림 2]는 이러한 동형암호기반 프라이버시 보존 머신러닝의 구조를 보여준다. 동형암호기반 머신러닝은 처리 속도가 일반 데이터의 수백 배 수준이라는 점과 AI 모델의 결과를 데이터를 암호화한 주체만 활용할 수 있다는 한계점이 있다.



[그림 2] 동형암호 기반 프라이버시 보존형 머신러닝 구조

차분 프라이버시(differential privacy)는 데이터에 노이즈를 추가하여 원본 데이터에 프라이버시를 제공하는 방법이다. 노이즈가 추가된 데이터를 사용해 AI가 학습하면 분류 정확도 등 AI의 성능이 크게 저하되리라 생각할 수 있지만, 실제로는 그렇게 많이 성능이 떨어지지 않는다. 또한 이 AI 모델에 학습데이터 복원을 위한 도치(inversion) 공격을 가할 경우 복원되는 데이터에도 노이즈가 추가되어 보호되는 특징도 있다. 연합학습(federated learning)은 앞에서 설명한 것처럼 학습데이터를 직접 전송하지 않고, 학습 결과인 파라미터만 전송하는 방법이다. 데이터를 직접 전송하지 않기 때문에 근본적인 프라이버시 보호가 될 것으로 생각할 수 있지만, 학습 결과 파라미터로부터 학습에 사용된 데이터를 복원하는 공격도 나와 있기 때문

에 근본적 해결이 될 수는 없다.

재현데이터(synthetic data)는 원본 데이터로 모델을 만들고 이 모델로부터 새로운 데이터를 생성하여 사용하는 방법이다. 새로 생성된 데이터를 재현데이터라고 부르며 원본 데이터 대신 재현데이터를 전송한다. 재현데이터도 원본이 아니므로 프라이버시가 보호된다고 생각할 수 있다. 하지만 재현데이터는 원본의 모델로부터 생성되므로 원본데이터와 같은 통계적 분포를 가지며, 이를 통해 원본 데이터가 갖는 민감 정보를 추론하는 것이 가능하다. 재현데이터의 프라이버시 위험도를 산정하는 연구는 별로 진행된 것이 없어 향후 연구가 필요하다.

4. 맺음말

AI에 대한 여러 가지 보안 공격들은 근본적인 해결 방안이 없는 창과 방패의 경주라고 할 수 있다. 그렇지만, AI 활용이 확대될수록 AI보안 문제는 더욱 큰 이슈가 될 것이기 때문에 지속적인 관심과 대비가 필요하다. AI 학습을 위해 사용되는 데이터의 프라이버시도 지속적으로 이슈가 될 것이다. 기존의 가명 및 익명처리 이외에도 비정형 데이터를 위한 프라이버시 보존형 머신러닝에 대한 집중적 연구가 필요한 실정이다.

[참고문헌]

[1] TTA, 개체명 태그 세트 및 태깅 말뭉치, TTAK.KO-10.0852

[2] 보건복지부, 보건의료 데이터 활용 가이드라인, 2021

[3] Ryu, G, et. al., Adversarial attacks by attaching noise markers on the face against deep face recognition, JOURNAL OF
INFORMATION SECURITY AND APPLICATIONS, 2021

[4] Na, H, et. al., Adversarial Attack Based on Perturbation of Contour Region to Evade Steganalysis-Based Detection, IEEE ACCESS, 2021

[5] Kwon, H, et. al., Classification score approach for detecting adversarial example in deep neural network, MULTIMEDIA TOOLS AND APPLICATIONS, 2021

[6] 김도완 외, Intrusion Detection System을 회피하고 Physical Attack을 하기 위한 GAN 기반 적대적 CAN 프레임 생성방법, 정보보호학회논문지, 2021

[7] 장진혁 외, 그래디언트 기반 재복원공격을 활용한 배치상황에서의 연합학습 프라이버시 침해연구, 정보보호학회논문지, 2021

[8] 류권상 외, 인공지능 보안 공격 및 대응방안 연구동향, 정보보호학회지, 2020

※ 출처: TTA 저널 제199호