Review Article

J Korean Soc Radiol 2019;80(2):259-273
https://doi.org/10.3348/jksr.2019.80.2.259
pISSN 1738-2637 / eISSN 2288-2928

대한영상의학회지

Exploiting the Vulnerability of Deep Learning-Based Artificial Intelligence Models in Medical Imaging: Adversarial Attacks

딥러닝 기반 의료 영상 인공지능 모델의 취약성: 적대적 공격

Hwiyoung Kim, PhD[1], Dae Chul Jung, MD[1,2], Byoung Wook Choi, MD[1,2]*

[1]Department of Radiology, Center for Clinical Imaging Data Science, Research Institute of Radiological Sciences, Yonsei University College of Medicine, Seoul, Korea
[2]Department of Radiology, Yonsei University Severance Hospital, Seoul, Korea

Received March 18, 2019
Accepted March 25, 2019

***Corresponding author**
Byoung Wook Choi, MD
Department of Radiology, Center for Clinical Imaging Data Science, Research Institute of Radiological Sciences, Yonsei University College of Medicine,
50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.

Tel 82-2-2228-7400
Fax 82-2-2227-8337
E-mail bchoi@yuhs.ac

ORCID iDs
Byoung Wook Choi
https://orcid.org/0000-0002-8873-5444
Hwiyoung Kim
https://orcid.org/0000-0001-7778-8973

Due to rapid developments in the deep learning model, artificial intelligence (AI) models are expected to enhance clinical diagnostic ability and work efficiency by assisting physicians. Therefore, many hospitals and private companies are competing to develop AI-based automatic diagnostic systems using medical images. In the near future, many deep learning-based automatic diagnostic systems would be used clinically. However, the possibility of adversarial attacks exploiting certain vulnerabilities of the deep learning algorithm is a major obstacle to deploying deep learning-based systems in clinical practice. In this paper, we will examine in detail the kinds of principles and methods of adversarial attacks that can be made to deep learning models dealing with medical images, the problems that can arise, and the preventive measures that can be taken against them.

Index terms Deep Learning; Artificial Intelligence; Medical Imaging

서론

지난 수년 동안 딥러닝(deep learning) 알고리듬(1)은 컴퓨터 영상처리 분야에 혁신을 가져왔다. 깊은 층위(layer)의 신경망을 이용한 깊은 신경망(Deep Neural Network) 모델들이 영상 객체분류(classification) 분야에서 기존 기계학습 모델들의 성능을 뛰어넘는 결과를 보이며 딥러닝이 본격적으로 주목받기 시작했다. 특히 구글의 딥러닝 기반의 얼굴 인식

(face recognition) 모델인 FaceNet (2)은 무려 99.96%의 인식률을 보여주었는데, 이는 인간의 얼굴 인식률이 평균적으로 97% 수준인 것을 감안했을 때 이를 뛰어넘는 성능을 보여준 것이다. 2016년에는 프로바둑기사 이세돌 9단과의 승부에서 구글의 알파고(AlphaGo) (3)가 승리하였다. 이는 얼굴인식과 같은 상대적으로 단순한 패턴인식(pattern recognition) 분야가 아닌, 인공지능이 절대로 인간의 능력을 뛰어넘을 수 없을 것 같았던 대표적 분야 중 하나인 바둑에서 인간을 이긴 것으로 평가되며 딥러닝 기술이 세간의 관심을 한 번에 끌어들일 만큼 큰 충격을 주었다.

급기야는 의학분야에서도 의학영상분야를 중심으로 딥러닝의 당찬 도전이 시작되었다. 안저영상(fundus image)을 이용한 당뇨성 망막변증(diabetic retinopathy) 진단에서 구글의 기계학습 기술이 안과 전문의와 동등하거나 앞서는 수준의 성능을 보여준 연구 결과가 발표된 것이 대표적이다(4). 이 당뇨성 망막변증 연구를 시작으로 다양한 의학분야에도 딥러닝의 흥미로운 응용 사례들이 연이어 발표되었다. 최근 영상의학(5-7), 병리학(8, 9), 안과학(4) 분야 등의 연구에서 의사들의 진단능과 동등성을 확보한 많은 사례가 있다. 경우에 따라 이러한 알고리듬의 성능은 개별 의사들의 진단능을 능가하는 것으로 발표되었다.

의료 영상처리에 기반한 컴퓨터 보조진단(Computer-Aided Detection and Diagnosis; 이하 CAD) 시스템의 목적은 주어진 의료 영상에서 병변(lesion)의 검출(detection) 또는 그 분류(classification) 결과를 제시하여 의사의 영상 판독 및 진단을 보조하는 것이다(10). 여느 기계학습 기반의 시스템들과 마찬가지로 CAD 시스템 또한 꾸준히 발전해왔다. 그간 공학자들과 의학자들의 노력으로 폐결절(lung nodule)과 유방암(breast cancer) 등의 질병들에 대하여 그 CAD 개발 사례와 성능이 보고되었으나, 특히 위양성(false-positive) 비율이 지나치게 높다는 등의 이유로 그 성능은 실제 임상 현장에서 쓰일 정도에 이르지 못하는 현실이었다(11-15).

그러나 앞서 언급하였듯이 기계학습 분야에 성능 혁신을 가져온 딥러닝 기술은 CAD 시스템에도 혁신을 가져오게 되었고, 결국에는 이전 CAD 시스템들의 성능을 뛰어넘게 되었다(4-9). 이러한 딥러닝 기반의 CAD 시스템에 대한 연구 성과들은 임상현장에 사용 가능한 수준으로 평가되어 활발히 상용화로도 이어지고 있다. 이미 미국 식품의약국(US Food and Drug Administration)과 식품의약품안전처(Ministry of Food and Drug Safety)에서는 딥러닝 기반의 CAD 시스템에 대한 의료기기 등록을 승인한 사례들이 보고되었다.

이러한 변화로 인하여 영상의학과나 병리학과 등의 의료 영상을 다루는 임상과를 중심으로 그 전문분야가 근본적으로 재구성될 필요가 있다는 주장이 제기되기도 했다. 일부 급진적인 연구자들은 영상의학 전문의를 포함하여라는 직업 자체가 곧 인공지능 기반의 시스템들로 대체될 것이라고 주장하기도 하였다.

그러나, 대표적으로 딥러닝 알고리듬에 대한 두 가지 단점이 지적되었다. 첫째는, 딥러닝 알고리듬이 내재적으로 가진 불확실성(uncertainty)이고(16), 둘째는 명시적으로 정의된 특징(hand-crafted feature) 기반의 기계학습 방법에 비해 학습된 딥러닝 모델은 블랙박스(black-box)와 같아 그 추론 결과가 설명가능(explainable) 하지 않다는 점이다(17). 위에 명시된 두 가지 단점은 특히 의학문제에 딥러닝 알고리듬을 적용하는 데에 큰 걸림돌이 된다.

본 종설에서는 특히 불확실성 문제에 의한 적대적 공격(adversarial attack) 문제에 대해 다룰

것이다. 의학영상을 다루는 딥러닝 모델들에 대해 어떠한 원리와 방식으로 적대적 공격이 이루어질 수 있으며, 이로 인하여 어떤 문제들이 발생할 수 있으며 적대적 공격을 차단할 수 있는 방법은 없는지 자세히 살펴보고자 한다.

적대적 공격의 정의와 예시

의학영상 문제에서의 딥러닝 학습 알고리듬의 이러한 진보와 더불어, 소위 '적대적 사례(adversarial example)'의 발견은 최첨단 학습 시스템에서도 예외 없이 취약성을 노출시켰다. Szegedy 등(18)은 딥러닝 모델이 다양한 컴퓨터 비전 분야에서 주목할 만한 정확성을 보이는 반면에, 이미지 분류 문제에서 흥미로운 약점을 보이는 것을 발견했다. 논문에서 그들은 딥러닝 모델의 높은 분류 정확도에도 불구하고, 사람의 눈으로는 거의 자각할 수 없는 수준의 미세한 교란 신호(perturbation)가 입력 영상에 추가된 것만으로도 영상의 분류 결과가 달라지는 등 놀라울 정도로 딥러닝 모델이 취약하다는 것을 보여주었다(Fig. 1).

이러한 적대적 공격은 딥러닝 분류 모델의 영상에 대한 분류 예측 결과를 완전히 바꿀 수 있고, 공격받은 모델은 잘못된 예측에 대하여 높은 신뢰도(confidence)를 보고하게 된다. 또한, 적대적 공격에 이용된 영상 노이즈를 이용하여 여러 네트워크 분류 모델을 동일한 메커니즘으로 속일 수 있다. Moosavi-Dezfooli 등(19)은 임의의 영상에 대하여 딥러닝 분류 모델을 속일 수 있는 '보편적인 노이즈(universal perturbations)'의 존재를 증명했다. Fig. 2의 예시들과 같이, 기존에 알려진 고성능의 영상 분류 딥러닝 학습 모델들이 원래는 높은 확신도(confidence)를 가지고 잘 분류하던 나무나 새 등의 원본 영상에 눈에 보이지 않는 pertubation을 더한 것만으로도 역시 높은 확신도를 가지고 잘못 분류하도록 할 수 있다.

이로써 적대적 사례와 같이 잘못된 분류를 유발하도록 설계된 입력을 통한 적대적 공격은 기계학습 커뮤니티에서 가장 인기 있는 연구 분야 중 하나가 되었다(20-23). 적대적 사례에 대한 많은 관심이 현재의 딥러닝 학습 모델의 한계를 밝혀내려는 노력에서 비롯된 것이지만, 실제로 딥러닝 학습 모델에 기반한 인공지능 프로그램들이 배포될 때 발생할 수 있는 사이버 보안 위협 때문에도 주목을 받는다.

학습모델의 일반화(Generalization)와 강인성(Robustness) 측면에서 바라본 적대적 공격

기계학습 모델의 성능평가는 그 모델의 일반화(generalization)의 정도를 확인하는 데에 중점

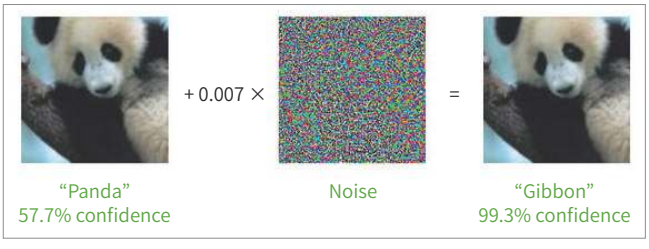
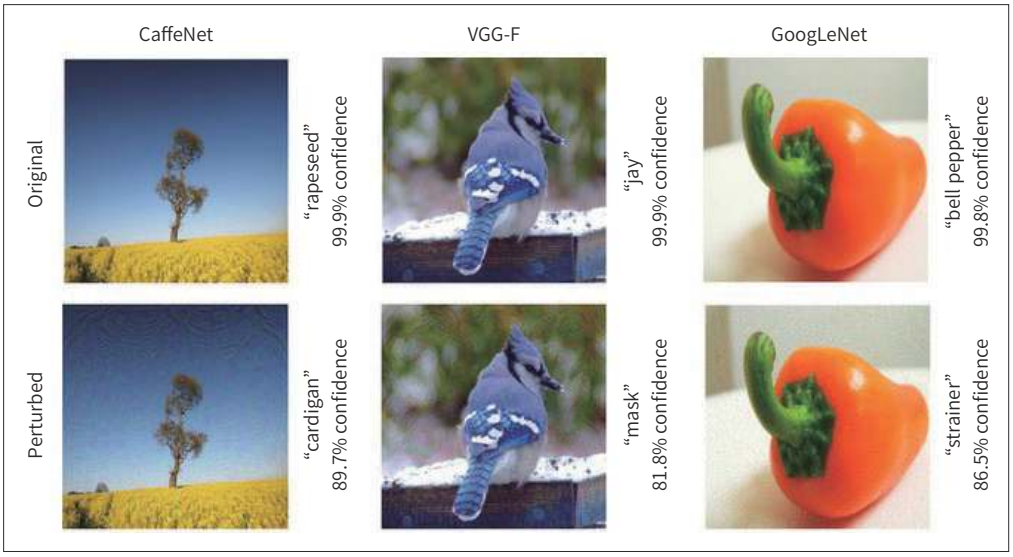


Fig. 1. An example of adversarial attack. A minimal perturbation added to an original image is able to cause a classifier to misclassify a panda as a gibbon. Adapted from Goodfellow et al. arXiv preprint 2014;arXiv:1412.6572, with permission of IEEE (28).

Fig. 2. Examples of adversarial attack on various deep learning models with "universal adversarial perturbations". Due to subtle perturbations added to the original images, the networks predicted wrong labels with high confidence. Adapted from Moosavi-Dezfooli et al. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE 2017:1765-1773, with permission of IEEE (19).



을 둔다. 이는 학습과정에서 쓰이지 않은 별도의 테스트 데이터에서의 성능에 대한 평가를 중심으로 확인할 수 있다. 예를 들어 의료 영상 데이터 세트와 같이 일반적으로 제한된 수의 학습 데이터를 가지는 경우 딥러닝 모델과 같이 대단위의 매개 변수를 갖는 모델을 학습에 이용하면 학습 데이터가 단순히 "암기(memorization)" 될 수 있기 때문에(이러한 경우, 학습과정에서 모델이 경험하지 못한 유형의 데이터에 대해서는 전혀 올바른 추론을 할 수 없게 된다) 학습 모델의 일반화에 대한 평가는 매우 중요하다.

위와 같이 별도의 테스트 데이터 세트를 이용한 학습모델의 일반화 평가 방식이 널리 사용되고 있지만, 데이터 세트의 이상치(outlier)나 불확실한 정답지(noisy label)에 대해 유연하게 대처하는 능력을 보는 학습모델의 강인성(robustness)을 평가하는 것은 별개의 문제이다.

즉, 기존의 학습모델 일반화 평가 방법과 이에 기반한 모델 학습과정은 학습모델이 과적합(overfitting) 되었는지에 대하여는 평가할 수 있지만 입력 데이터의 변화에 대한 모델의 민감도에 대해서는 평가할 수 없다. 적대적 공격은 이러한 부분을 파고들어 딥러닝 학습모델의 허점을 드러내게 한다.

학습모델의 강인성에 대한 평가는 모델의 성능이 한계점까지 도달했을 때 잠재적 오류에 대한 확률(불확실성)을 추정하는 방식으로 이루어져야 한다. 이러한 모델의 강인성 평가 방법 및 이에 기반하여 강인한 학습모델을 구현하는 방법론은 지금도 많은 컴퓨터 공학자들이 도전하고 있는 연구 주제이며, 표준적인 방법론은 아직 존재하지 않는다. 이러한 상황에서 학습모델에 대한 적대적 공격이 가능하다는 사실은 딥러닝 기반의 학습모델들의 취약성으로 남을 수밖에 없기 때문에 각별한 주의가 필요하다(24).

적대적 공격의 유형과 적대적 사례의 생성방법

딥러닝 학습모델에 대한 대부분의 적대적 공격은 적대적 사례를 통해 수행된다. 공격자의 목표에 따라 적대적 공격은 표적(targeted) 공격과 비표적(non-targeted) 공격으로 나눌 수 있다. 공격자가 분류기의 출력을 사전에 특정한 대상 레이블로 의도적으로 변경하려고 하는 경우에 이 공격을 표적 공격이라 한다. 비표적 공격의 경우, 공격자의 목적은 단지 분류자가(무엇이 되었든) 잘못된 레이블을 선택하도록 하는 것이다. 직관적으로 알 수 있듯이, 일반적으로 비표적 공격이 표적 공격에 비하여 성공률이 높다(25).

또한 Papernot 등(26)에 따르면, 모델 학습의 워크 플로우 상에서 어떤 단계에 적대적 공격이 반영되는가에 따라서 두 가지 유형의 공격이 있다. 적대적 사례가 모델 학습 단계에 적용되어 훈련 중 모델을 방해하려고 시도하는 경우, 이러한 공격은 중독(poisoning) 공격이라 한다. 예를 들어 학습 데이터셋에 학습모델이 오류를 일으킬 수 있는 예제들을 고의로 추가하는 방식도 중독 공격이 될 수 있다. 반면에, 적대적 사례는 모델의 추론(inference) 단계에서도 고의적으로 모델이 오작동을 일으키도록 사용될 수 있으며, 이러한 공격을 회피(evasion) 공격이라 한다.

그리고, 딥러닝 학습모델에 대한 공격 시나리오는 공격자가 공격의 대상이 되는 모델에 대해 가진 정보의 양에 따라 다를 수 있다. 공격자가 공격의 대상이 되는 모델에 대한 완전한 구조(모델의 아키텍쳐, 파라미터 및 하이퍼 파라미터 등) 및 때로는 학습 데이터를 알고 있는 상황에서 적대적 공격을 수행하는 화이트박스(white box) 공격과 공격자가 모델에 대한 내부구조를 파악하지 못한 상태에서 수행하는 블랙박스(black box) 공격으로 나뉘게 된다. 블랙박스 공격의 경우에는 공격자가 확보할 수 있는 정보는 특정 입력에 대한 추론결과(predicted label) 뿐이다(25).

위의 언급된 적대적 공격의 유형들에 대해 직관적으로, 학습모델에 대한 화이트박스 공격이면서 학습 데이터셋에 대해 적대적 사례를 만들어 공격할 수 있는 시나리오가 가장 공격 성공률이 높을 것이다. Bun과 Steinke(27)는 이를 완전한 정보(perfect-knowledge)에 기반한 공격이라 지칭하고 있다. 하지만 이러한 공격 시나리오는 현실적이지 않으며, 대부분의 현실적인 공격 시나리오는 제한된 정보(limited-knowledge)에 기반한 블랙박스 공격이 될 것이다(27).

방법론적으로 현재까지 알려진 가장 보편적인 적대적 공격은 기울기(gradient) 기반 방법이다. 즉, 공격자는 입력 영상에 대하여 모델의 손실 함수(loss function)의 기울기 방향으로 영상을 수정한다. 아래에서는 적대적 공격을 수행하는 몇 가지 주요 알고리듬에 대해 기술하고자 한다. Table 1은 현재까지 발표된 적대적 공격과 그 방어법에 관한 문헌들을 정리한 것이다.

Fast Gradient Sign Method (FGSM)

Goodfellow 등(28)에 의해 2014년 제안된 알고리듬으로 적대적 공격을 위한 샘플을 만드는데 가장 간단하며, 계산적으로 효율적인 방법이다. 관련한 연구에서 벤치마킹을 위한 baseline 모델로서 주로 사용된다. 학습 시 기울기 하강(gradient-descent) 과정에서 사용된 기울기 방향과 반대 방향에 해당하는 perturbation을 영상에 추가하게 되면 학습을 저하시키는 효과를 얻을 수 있을 것이라는 매우 직관적이고 간단한 아이디어에서 출발한 알고리듬이다. 즉, Fast Gradient Sign

Table 1. Adversarial Attack and Defense Methods

	References
Attack	
White box	FGSM, Goodfellow et al. (28), L-BFGS, Szegedy et al. (18), BIM & ILCM, Kurakin et al. (22), JSMA, Papernot et al. (30), C&W attacks, Carlini et al. (33), DeepFool, Moosavi-Dezfooli et al. (29), Universal Perturbations, Moosavi-Dezfooli et al. (19), NewtonFool, Jang et al. (31)
Black box	One-pixel, Su et al. (54), UPSET & ANGRI, Sarkar et al. (55), Substitute Model Training, Papernot et al. (56)
Defense	
Gradient masking	Buckman et al. (51), Papernot et al. (20), Zantedeschi et al. (52), Lyu et al. (45), Nguyen et al. (46)
Adversarial training	Szegedy et al. (18), Goodfellow et al. (28), Sun et al. (53)

BIM & ILCM = Basic iterative method and iterative least-likely class method, C&W = Carlini and Wagner, FGSM = Fast Gradient Sign Method, JSMA = Jacobian-based Saliency Map Attack, L-BFGS = Limited-memory Broyden–Fletcher–Goldfarb–Shanno, UPSET and ANGRI = Universal Perturbations for Steering to Exact Targets and Antagonistic Network for Generating Rogue Images

Method (이하 FGSM) 방법은 손실함수를 증가시키는 perturbation을 영상에 추가하여 적대적 공격을 수행하는 방법이다. 알고리듬의 단순성으로 인해 공격 성공률은 낮은 편이다(ImageNet 데이터로 학습된 네트워크에 대해 63~69%).

$$X^{adv} = X + \varepsilon sign\,[\nabla x J(X, y_{true})]$$

One-Step Target Class Methods

FGSM의 대체 알고리듬 중의 하나로, 표적공격을 하므로 Targeted FGSM 방법으로도 불린다. 한 특정 레이블 y_{target}에 속할 가능성이 적은 어떠한 이미지 X에 대해서 그 이미지가 특정 레이블에 속할 확률 $p\,(y_{target} \mid X)$을 최대화하는 방향으로 perturbation을 추가하는 학습을 진행하게 된다. 유사하지 않거나 전혀 연관성이 없어 보이는 레이블로 오인하도록 표적공격을 수행할 수 있으므로, 적대적 공격에 의한 효과를 최대한으로 할 수 있는 알고리듬이다(22).

$$X^{adv} = X - \varepsilon sign\,[\nabla x J(X, y_{target})]$$

DeepFool

Moosavi-Dezfooli 등(29)이 2016년 제안한 알고리듬으로 탐욕 알고리즘이기 때문에 다소 느리지만 Jacobian-Based Saliency Map Attack 알고리즘보다는 월등히 빠르고, FGSM 알고리즘보다 정확하기 때문에 유용하게 활용된다. 특정 입력 포인트에서 손실함수를 반복적으로 선형화하고(해당 입력 포인트에서 손실함수에 접하는 접선벡터를 구하여 근사한다), 이 선형 근사가 정확하다면 학습모델의 추론 결과를 전환하는 데 필요한 최소한의 perturbation이 적용되는 것이다.

선형 대수적으로 이때 "최소한"의 perturbation은 입력벡터를 결정경계(decision boundary)로 투영한 벡터의 크기로 정의할 수 있다.

Jacobian-Based Saliency Map Attack (JSMA)

Papernot 등(30)이 표적공격의 하나로 제안한 알고리듬이다. 딥러닝 학습모델이 입력 영상에 대해 분류 결과를 판단하는 데 있어 입력 영상에서 어느 특정 부분들이 주요하게 영향을 미쳤는지를 나타낸 일종의 주의맵(attention map)인 saliency map을 gradient 기반으로 구성하여 이를 토대로 입력 영상에서 어느 부분들에 대해 최소한의 수정을 가하면 원하는 레이블로 분류 모델이 오작동하도록 하는 것이 가능할지를 판단하는 방법이다.

NewtonFool

Jang 등(31)이 제안한 방법으로서, 입력 영상에서 화소 단위로 특정 레이블에 높은 확률로 속하는 화소와, 반대로 해당 레이블에 높은 확률로 속하지 않는 화소를 뉴턴 알고리즘을 통해 찾는 방법이다. 이 알고리즘도 학습모델의 추론 결과를 전환하는 데 필요한 최소한의 perturbation을 탐색하도록 작동한다.

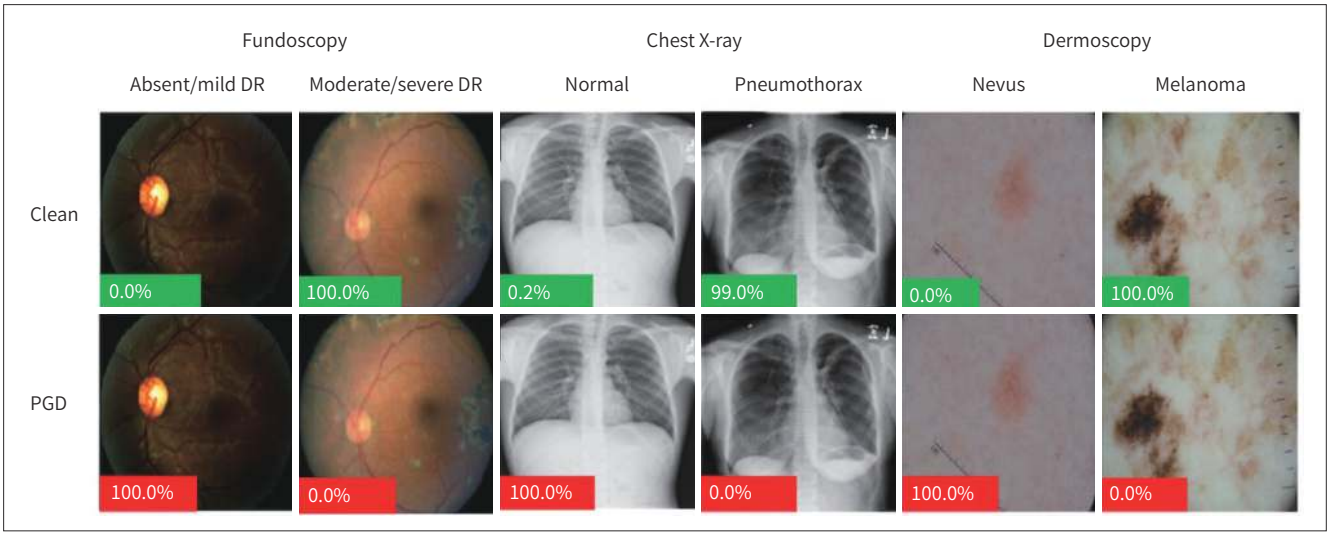
의료 문제에서의 적대적 공격

딥러닝 학습모델의 성능을 교란시킬 수 있는 적대적 사례가 여러 가지 방법으로 생성될 수 있음을 이론적으로 확인하였다면, 그다음 단계는 실제로 학습된 딥러닝 모델이 어느 정도로, 어떤 유형으로 적대적 공격이 가능한지 실험해 보는 것이다. 많은 연구자들이 자율주행 자동차와 같은 특정 실제 시스템에 대한 적대적 공격의 실현 가능성과 가능한 원인에 대해 논의하였다(32-34). 그러나 여전히 의학 문제에 적용된 딥러닝 학습 시스템에 대한 적대적 공격의 가능성은 충분히 논의되지 못한 상태이다.

Taghanaki 등(35)은 Chest X-ray 14 데이터셋을 이용한 흉부 X선 영상에 집중하여 10가지의 적대적 공격 방법을 통한 딥러닝 학습모델의 취약성을 광범위하게 테스트했다. 흉부 X선 영상에 적용된 기울기 기반의 적대적 공격들이 매우 성공적이라는 것을 보여주었다. 반면에 화소 단위로 영상의 변조를 시도하는 알고리듬으로는 공격에 성공하지 못했다(일반적인 RGB 영상에는 작동하는 것으로 알려져 있다). 또한 CNN 모델의 풀링(pooling) 과정에 대한 일부 적대적 공격 방법에 대해 유의미한 성능 저하를 확인했으며, 심지어 특정 클래스에 대한 공격은 완전히 실패하도록 방어가 가능하다는 결과도 보여주었다.

하버드의대의 Finlayson 등(36)은 흉부 X선 영상, 안저 영상, 피부 영상 등 3가지의 대표적인 의료 영상 분석에 딥러닝이 활용된 사례에 대해 적대적 공격을 시도하여 성공한 결과를 제시하였다. 흉부 X선 영상을 이용한 기흉(pneunothorax), 안저 영상을 이용한 중증도 이상의 당뇨성망막변증(diabetic retinopathy), 피부 영상을 이용한 악성흑색종(melanoma) 등에 대한 각각의 자동 진단 학습 네트워크를 구성하여 실험하였다. 정상적인 데이터셋에서 area under a receiver operat-

Fig. 3. Examples of quantitative results of an adversarial attack on medical image (36). Each original image (upper) well-classified with high confidence (green = model is correct) for the given diagnosis was misclassified after noise added to its corresponding original image (lower) with high confidence (red = model is incorrect). Adapted from Finlayson et al. arXiv preprint 2018;arXiv:1804.05296, with permission of IEEE (36). DR = diabetic retinopathy



ing characteristic curve (이하 AUROC) 기준 0.90 내외의 성능을 보이던 학습모델들이 적대적 공격에 의해 AUROC 0.10 수준으로 정상적인 진단 성능을 보여주지 못하는 것을 확인했다(Fig. 3). 또한 그러한 적대적 공격이 현실적으로 어떻게 수행될 수 있는지에 대한 구체적인 예를 제공하였다. 논문에서 그들은 의학 분야가 금전적 인센티브 및 기술적 취약성 측면에서 딥러닝 학습모델에 대한 적대적 공격에 특별히 취약할 수 있다고 주장하였다. Finlayson 등(36)이 기술한 바에 의하면 적대적 사례가 실제 의료환경에서 악용될 수 있는 시나리오는 다음과 같다.

의료보험의 부정수급

특히 행위별수가제(fee-for-service)인 경우에 부정수급을 노린 적대적 공격의 가능성이 있다. 미래에 보험 회사는 특정 의료행위가 필요한 것이었는지 확인하기 위하여 의료 영상을 입력으로 받는 딥러닝 자동판독 시스템을 운영할 수 있다. 이때, 제출된 의료 영상에 보험 회사의 시스템을 속일 수 있는 적대적 사례를 포함하여 실제로 질병이 없는 경우이지만 의료보험의 부정수급을 위한 용도로 악용될 수 있다.

약물 임상시험 대상자 선별을 위한 적대적 사례 사용

승인 후 4년 동안 항암제의 매출은 16억 7000만 달러에 달한다. 개발된 신약의 임상적 효능을 제대로 검증하기 위해서 임상시험 참가자에 대해 임삼시험 대상으로 적절했는지에 대한 평가도 함께 이루어지는데, 이때 환자의 병기가 임상시험의 참여 조건에 적절하도록 변조된 적대적 사례를 포함한 의료 영상 자료를 제출하여 이러한 검증시도를 회피할 가능성이 있다.

의학영상 분석 모델에 대한 적대적 공격 외에도 다른 의료 분야에 적용 가능한 시나리오들이 제시되기도 하였다. Papangelou 등(37)은 임상시험 참가자 선별을 위한 학습모델을 속일 수 있는

소위 적대적 환자(adversarial patient)의 개념을 제시하고, 여러 가지 유형의 적대적인 사례 생성 알고리듬이 적대적 환자를 생성할 수 있는 가능성을 확인하였다.

물론 이러한 적대적 사례를 역으로 활용하여 딥러닝 학습모델의 강인성을 평가할 수도 있다. Paschali 등(24)은 다양한 의료 영상 분석을 위한 딥러닝 학습모델의 강인성을 평가하는 척도로서 적대적 사례를 활용하였다. 기존의 딥러닝 학습모델의 성능에 대한 평가는 별도의 테스트 데이터셋에 대한 분류 정확도 등을 보는 것으로 이루어져왔다. 이 연구에서는 딥러닝 학습모델들이 노이즈가 있거나, 이상치(outlier)에 가까운 극단적인 경우의 입력 영상들에 대하여도 좋은 성능이 확보되는지 확인하기 위하여 적대적 사례들을 활용한 평가 방법을 제시하였다. 영상 분류(classification)와 분할(segmentation) 문제 각각에 대해 좋은 성능을 보이는 Inception (38) 및 U-Net (39)과 같은 딥러닝 학습 네트워크 모델로 피부 병변의 분류 및 뇌 MRI 영상에서의 분할 문제에 대해 학습시킨 후 광범위한 실험을 통하여 모델의 강인성은 테스트 데이터셋의 구성에 따라 큰 차이를 보임을 확인하였고, 이 평가 과정에서 적대적 사례의 활용 가능성을 제시하였다.

위 연구에서는 부수적으로 분할 문제를 해결하는 데 있어 U-Net과 같이 밀집블록(dense block)과 통과연결(skip connection) 구조를 갖는 학습 모델의 경우, 모델의 강인성도 확보되는 학습이 가능함을 확인하였다. 또한 깊은 층위의 모델을 이용할 경우 적대적 공격에 대한 저항력을 확보할 수 있음 또한 확인하였다.

또한 Sun 등(40)은 의료 영상이 아닌 의무기록을 이용한 응급환자 선별과제에 적대적 공격을 재밌는 방법으로 응용하였다. 의무기록과 같은 시계열 데이터에 대하여 순환 신경망(Recurrent Neural Network) (41)이나 장-단기 기억 신경망(Long-Short Term Memory) (42)과 같은 딥러닝 학습모델을 이용하여 일련의 의무기록의 변화에서 패턴을 감지하여 환자의 상태를 예측한다든가(예: 특정 응급질환의 발생 여부) 응급환자를 선별하는 데에 활용할 수 있다. Sun 등(40)은 연구를 통하여 이러한 일련의 의무기록에서 어떠한 특정 의무기록들이 어떠한 조건에서 응급환자 선별에 민감하게 영향을 미치는지를 적대적 공격을 통하여 확인하였다. 적대적 공격에 취약하게 작용한 의무기록은 역으로 환자의 응급환자 선별에 주요하게 작용하는 하나의 팩터로서 이해할 수 있다는 것이다(Fig. 4)

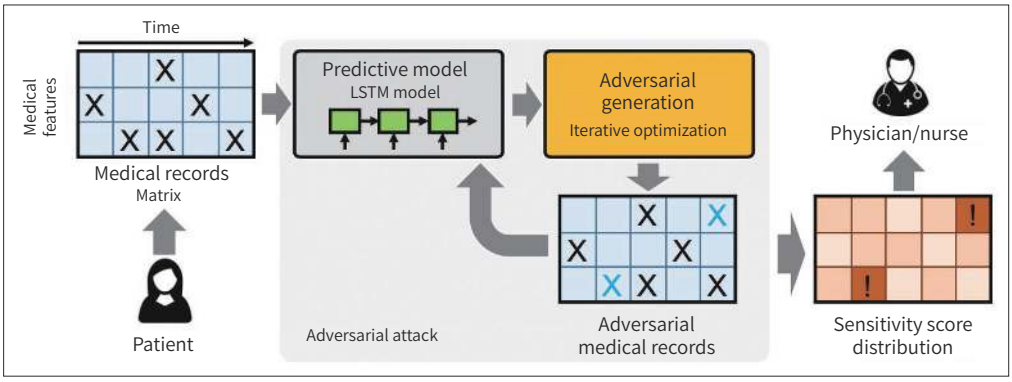
의료문제에서 딥러닝 모델이 갖는 취약성의 근원들

Finlayson 등(36)은 특히 의료문제에 적용된 딥러닝 모델들이 갖는 취약성에 대해서 분석하면서, 딥러닝 모델 자체의 강인성 측면 외에도 의료 환경 또는 데이터가 갖는 내재적 문제들로 인해 적대적 공격 등에 대한 취약성이 발생한다는 주장을 하였다.

의학은 내재적으로 불확실성을 가진다

개나 고양이 분류 문제와 같은 대부분의 일반적인 영상 분류 작업과 비교하여, 의학 영상의 소위 정답지는 논란의 여지가 있는 경우가 많으며 영상의학 전문의조차도 영상소견을 판단하는 데 있어 의견이 맞지 않을 수 있다(11, 32, 42). 따라서 의학 영상을 자동 진단하는 딥러닝 학습모델의

Fig. 4. A schematic representation of the proposed algorithm for identifying vulnerable locations in electronic medical records (40). Medical records generated by adversarial attacks can be used to derive a susceptibility score distribution of the medical records. Adapted from Sun et al. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM 2018:855-868, with permission of ACM (40).



학습 자체가 모호하게 이루어질 수 있으며, 실제로 판독이 어려운 케이스의 영상들 만을 선택적으로 추가하여 적대적 공격을 통해 모델의 학습과정을 교란 시킬 수 있는 여지가 크다.

의료 영상은 표준화되어 있다

예를 들어, 자율주행을 위해 자동차에 설치된 카메라로 받는 일반적인 영상들과 달리 의료 영상은 특정 부위를 촬영하기 위한 프로토콜이 표준화되어 있어 동일한 해부학적 구조가 여러 환자들의 영상에서 크게 다르지 않은 위치와 영상 품질로 나타나게 된다. 반면에 자율주행의 경우, 영상에서 확인하고자 하는 다른 자동차, 표지판 등이 영상 내에서 다양한 위치에 나타날 수 있고, 날씨나 일과 중 시간, 햇빛과 같은 조명의 위치에 따라 천차만별의 영상 품질을 갖게 된다. 이러한 점이 의료 영상이 더욱 일반 영상들에 비해 상대적으로 적대적 공격에 취약하도록 하는 원인 중 하나이다.

주로 사용되는 딥러닝 네트워크 구조가 정해져 있다

의료 영상의 자동 진단 문제에 대해 그동안 발표된 연구들은 거의 동일하거나 유사한 딥러닝 네트워크 구조를 이용하였다. 대부분 특정 작업에 맞게 미세 조정되었을 뿐 일반적인 영상 분류 문제에서 좋은 성적을 거둔 딥러닝 네트워크 구조를 이용하였고, 특히 문제에 따라 충분한 수의 환자 데이터 확보가 어려운 경우 전이학습(transfer learning)을 위해 일련의 미리 학습된 딥러닝 모델을 이용하는 경우가 많다. 사용된 딥러닝 학습모델 구조의 다양성이 부족하다는 점은 공격자로 하여금 그 구조를 미리 파악하고 수행하는 화이트박스 공격이 가능하게 하여 상대적으로 높은 확률로 공격을 성공시킬 수 있는 여지를 제공한다.

학습에 이용된 의료 데이터를 공개하도록 한다

학습에 이용된 데이터를 공개하도록 하는 요구는 연구 성과의 원활한 공유와 공개 논의가 가능하도록 하여 연구 생태계 내에서 연구의 발전을 빠르게 하는 긍정적인 효과가 있지만, 학습 데이터가 공개된다는 것은 공격자에 입장에서는 적대적 공격의 성공 확률을 높일 수 있는 좋은 힌트가 될 수 있다.

병원 인프라는 업데이트하기가 매우 어렵다

의료용 소프트웨어는 의료기기 승인 문제 등 여러 문제와 맞물려 업데이트 및 수정 작업에 많은 시간과 비용이 소요된다. 따라서 의료 영상의 자동 진단 모델과 같은 의료용 소프트웨어에 취약성이 존재하더라도 병원 인프라 업데이트에 소요되는 시간과 비용 문제로 인해 발빠른 대처가 어려울 수 있다. 또한 이러한 문제에 대처할 수 있는 전문 인력을 병원 내에 확보하기 어려운 여건이라는 점도 고려해 보아야 할 점이다.

의료 영상 데이터는 개인 고유의 특성을 갖는다

굳이 적대적 공격을 수행하지 않더라도, 일반적인 경우에는 단순히 테스트 데이터셋 중의 특정 영상을 공격자가 원하는 레이블로 분류되는 영상으로 대체하는 것만으로도 자동 분류 모델에 대한 공격 목적을 달성할 수 있다. 하지만 안저 영상이나 X선 영상과 같은 의료 영상은 흔히 지문과 같이 각 개개인의 고유한 특성이 드러나는 영상 자료이므로, 동일한 환자가 이전에 촬영했던 영상과 입력 영상을 비교하여 신원이 일치하는지 확인하는 알고리듬을 설계하여 단순한 영상 대체 공격에 대해 방어가 가능하다. 하지만 신원이 확인 가능한 선에서 최소한의 영상 perturbation 만으로 이루어지는 적대적 공격에는 여전히 취약할 수밖에 없다.

적대적 공격에 대한 방어법

딥러닝 학습 모델을 적대적 공격에 보다 견고하게 하기 위해서 최근에 여러 가지 방어 알고리듬이 제안되었다. 적대적 공격에 대한 방어는 크게 세 가지 접근 방식으로 나누어 볼 수 있다.

적대적 훈련(Adversarial Training)

가장 직관적으로 쉽게 떠올려볼 수 있는 알고리듬이다. 자동 진단 모델을 학습시킬 때, 적대적 사례로서 작동할 수 있는 모든 경우의 수를 미리 학습 데이터셋에 포함시키는 것이다. Szegedy 등(18)은 이러한 적대적 훈련에 대한 개념과 그 방법론을 제안하였다. 이 연구에서 가능한 적대적 사례들에 대해 모델이 충분히 학습되었다는 가정하에 적어도 사전에 학습했던 perturbation의 패턴들에 대해서는 모델이 강인성을 갖는다고 주장하였다. 부수적으로 이러한 적대적 훈련 방법이 모델의 정규화(regularization)에 도움이 되고(28) 과적합(over-fitting)을 방지하기도 한다(43)는 연구 결과들이 존재한다. 실제로 이런 원리로 인해 보통은 학습데이터 개수의 부족함을 극복하기 위해 사용하는 데이터 증강(data augmentation) 적용시에 영상에 다양한 노이즈를 추가하는 것이 학습모델의 성능 향상에 도움을 주기도 한다(44).

Gradient Masking/Distillation

적대적 훈련과 같이 학습모델을 전혀 건드리지 않은 채 학습 데이터와 학습 과정에만 변화를 주는 방법은 학습 과정을 더 복잡하게 하거나 부담을 주게 되어 직관적이긴 하나 실용적인 방법은 아니다. 특히, 충분한 수와 다양성이 보장된 적대적 사례를 생성하는 과정 없이는 적대적 훈련은

그 성능을 보장하기 어렵다. 대부분의 적대적 공격은 모델의 추론 과정에서의 gradient를 관찰함으로써 이루어진다는 점에 착안하여, 학습모델의 gradient가 출력으로서 그대로 노출되는 것을 방지하거나(gradient masking), 학습모델의 구조상 gradient 자체를 일종의 정규화 방법과 같이 두드러지지 않게 하여 적대적 공격의 학습 방향에 힌트를 주지 않도록 하는 방법(distillation) (20)들이 제안되었다. Lyu 등(45)은 학습 과정에서 과한 gradient에 패널티를 주는 일종의 정규화 방식으로 학습모델의 강인성을 확보하였다. 더 나아가 Nguyen 등(46)은 학습모델의 출력에 노이즈를 추가하는 방식의 gradient masking 방법을 제안하였다.

Feature Squeezing

그 외에 적대적 공격을 막기 위한 방법으로는 본래의 학습모델과 별도로, 주어진 입력에 대해 적대적 사례인지 아닌지를 판단하는 학습모델을 덧붙이는 것이 제안되기도 했다. Feature Squeezing은 그 방법 중 하나로 Xu 등(47)이 제안하였다. 그들은 딥러닝 학습모델에 별도로 다음과 같은 두 개의 기능을 추가하였다. 1) 영상의 인코딩을 단순화하여 표현 색상의 깊이(depth)를 축소하고, 2) 영상에 대해 공간적 평활화(smoothing) 필터를 적용하였다. 이로써 주어진 원본 영상과 위의 과정으로 표현이 압축된 영상에 대한 학습 네트워크의 추론 결과를 비교하여 두 결과 간에 큰 차이가 발견되면 주어진 영상이 적대적인 예시라고 간주하는 방식이다. Feinman 등(48)은 학습모델의 특징 공간(feature space)에서의 불확실성(uncertainty)의 평가(특히, dropout을 이용함으로 인한) 및 밀도추정(density estimation)을 수행함으로써 주어진 영상에서의 perturbation을 감지하여 적대적 사례 여부를 판별하는 방법론을 제안하였다(48, 49).

그 외에 다수의 학습모델을 앙상블(ensemble)하여 시스템을 구성하면, 특정 모델에 대한 화이트박스 공격을 피할 수 있음은 물론이고, 다수의 학습모델에 범용적으로 적용되는 적대적 공격은 개발이 어렵다는 점으로 인해 좋은 방어법이 된다는 연구 결과도 있다(50). Strauss 등(50)은 Modeified National Institute of Standards and Technology와 Canadian Institute For Advanced Research-10 데이터 세트에 대해 앙상블 방법이 학습모델의 영상 분류 정확도를 향상시킬 뿐 아니라 적대적 공격에 대한 강인성을 증가시킨다는 것을 실험으로 증명하였다.

결론

본 종설에서는 딥러닝 알고리듬의 불확실성에 의해 의학 영상을 다루는 자동 진단 시스템에서 발생할 수 있는 잠재적인 취약성 문제에 대해 다루어 보았다. 딥러닝 학습모델 성능의 비약적인 발전으로 인해 영상의학을 중심으로 하여 기계학습 모델들이 실제 임상현장에서 의사를 보조하여 진단능을 높이고 작업의 효율성을 증대시켜줄 것이라는 기대가 많다. 실제로 이러한 기대로 인해 많은 병원과 민간 기업 등에서 의학 영상을 이용한 자동 진단 학습모델 개발 경쟁이 뜨겁다. 앞으로는 많은 딥러닝 기반의 자동 진단 프로그램들이 의료 환경에서 사용될 것이다. 그러나 본 종설에서 지적한 바와 같이 딥러닝 기반의 의료 영상 자동 진단 모델들은 적대적 공격에 취약하므로, 몇몇 연구의 예시로서 제시된 가능한 시나리오들에 대비한 연구가 동반되어야 할 것이다.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

REFERENCES

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521;436-444
2. Schroff F, Kalenichenko D, Philbin J. *FaceNet: a unified embedding for face recognition and clustering.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE 2015: 815-823
3. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. *Nature* 2017;550:354-359
4. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316:2402-2410
5. Liu F, Zhou Z, Samsonov A, Blankenbaker D, Larison W, Kanarek A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology* 2018; 289:160-169
6. McBee MP, Awan OA, Colucci AT, Ghobadi CW, Kadom N, Kansagra AP, et al. Deep learning in radiology. *Acad Radiol* 2018;25:1472-1480
7. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint* 2017;arXiv:1711.05225
8. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199-2210
9. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast cancer. *JAMA* 2017; 318:2184-2186
10. Giger ML, Suzuki K. *Computer-aided diagnosis. In Biomedical Information Technology.* New York: Academic Press 2008
11. Böröczky L, Zhao L, Lee KP. Feature subset selection for improving the performance of false positive reduction in lung nodule CAD. *IEEE Trans Inf Technol Biomed* 2006;10:504-511
12. Tan M, Deklerck R, Jansen B, Bister M, Cornelis J. A novel computer-aided lung nodule detection system for CT images. *Med Phys* 2011;38:5630-5645
13. Cao P, Liu X, Yang J, Zhao D, Li W, Huang M, et al. A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules. *Pattern Recognit* 2017; 64:327-346
14. Baker JA, Rosen EL, Lo JY, Gimenez EI, Walsh R, Soo MS. Computer-aided detection (CAD) in screening mammography: sensitivity of commercial CAD systems for detecting architectural distortion. *AJR Am J Roentgenol* 2003;181:1083-1088
15. Dromain C, Boyer B, Ferré R, Canale S, Delaloge S, Balleyguier C. Computed-aided diagnosis (CAD) in the detection of breast cancer. *Eur J Radiol* 2013;82:417-423
16. Gal Y. *Uncertainty in deep learning.* Los Altos: IEEE/ACM Transactions on Audio, Speech, and Language Processing 2017
17. Holzinger A. *From machine learning to explainable AI.* In 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA). Piscataway: IEEE 2018:55-66
18. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. *arXiv preprint* 2013;arXiv:1312.6199
19. Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. *Universal adversarial perturbations.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE 2017:1765-1773
20. Papernot N, McDaniel P, Wu X, Jha S, Swami A. *Distillation as a defense to adversarial perturbations against deep neural networks.* In 2016 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE 2016:582-597
21. Melis M, Demontis A, Biggio B, Brown G, Fumera G, Roli F. *Is deep learning safe for robot vision? adversarial examples against the icub humanoid.* In Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE 2017:751-759

22. Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. *arXiv preprint* 2016;arXiv:1607.02533
23. Athalye A, Engstrom L, Ilyas A, Kwok K. Synthesizing robust adversarial examples. *arXiv preprint* 2017;arXiv:1707.07397
24. Paschali M, Conjeti S, Navarro F, Navab N. Generalizability vs. robustness: adversarial examples for medical imaging. *arXiv preprint* 2018;arXiv:1804.00504
25. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access* 2018;6:14410-14430
26. Papernot N, McDaniel P, Sinha A, Wellman M. Towards the science of security and privacy in machine learning. *arXiv preprint* 2016;arXiv:1611.03814
27. Bun M, Steinke T. *Concentrated differential privacy: simplifications, extensions, and lower bounds.* In Theory of Cryptography Conference. Berlin, Heidelberg: Springer 2016:635-658
28. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv preprint* 2014; arXiv:1412.6572
29. Moosavi-Dezfooli SM, Fawzi A, Frossard P. *Deepfool: a simple and accurate method to fool deep neural networks.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE 2016;2574-2582
30. Papernot N, Mcdaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. *The limitations of deep learning in adversarial settings.* In 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE 2016:372-387
31. Jang U, Wu X, Jha S. *Objective metrics and gradient descent algorithms for adversarial examples in machine learning.* In Proceedings of the 33rd Annual Computer Security Applications Conference. Piscataway: IEEE 2017:262-277
32. Evtimov I, Eykholt K, Fernandes E, Kohno T, Li B, Prakash A, et al. Robust physical-world attacks on deep learning models. *arXiv preprint* 2017;arXiv:1707.08945
33. Carlini N, Wagner D. *Towards evaluating the robustness of neural networks.* In 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE 2017
34. Lu J, Sibai H, Fabry E, Forsyth D. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint* 2017;arXiv:1707.03501
35. Taghanaki SA, Das A, Hamarneh G. *Vulnerability analysis of chest x-ray image classification against adversarial attacks.* In Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Cham: Springer 2018
36. Finlayson SG, Chung HW, Kohane IS, Beam AL. Adversarial attacks against medical deep learning systems. *arXiv preprint* 2018;arXiv:1804.05296
37. Papangelou K, Sechidis K, Weatherall J, Brown G. *Toward an understanding of adversarial examples in clinical trials.* In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer 2018:35-51
38. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. *Going deeper with convolutions.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE 2015
39. Ronneberger O, Fischer P, Brox T. *U-net: convolutional networks for biomedical image segmentation.* In International Conference on Medical image computing and computer-assisted intervention. Cham: Springer 2015;234-241
40. Sun M, Tang F, Yi J, Wang F, Zhou J. *Identify susceptible locations in medical records via adversarial attacks on deep predictive models.* In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM 2018:793-801
41. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans Pattern Anal Mach Intell* 2009;31:855-868
42. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput* 2000;12:2451-2471
43. Miyato T, Dai AM, Goodfellow I. Adversarial training methods for semi-supervised text classification. *arXiv preprint* 2016;arXiv:1605.07725
44. Zheng S, Song Y, Leung T, Goodfellow I. Improving the robustness of deep neural networks via stability training. *arXiv preprint* 2016;arXiv:1604.04326

45. Lyu C, Huang K, Liang HN. *A unified gradient regularization family for adversarial examples.* In 2015 IEEE International Conference on Data Mining. Piscataway: IEEE 2015:301-309
46. Nguyen L, Wang S, Sinha A. *A learning and masking approach to secure learning.* In International Conference on Decision and Game Theory for Security. Cham: Springer 2018:453-464
47. Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks. *arXiv preprint* 2017;arXiv:1704.01155
48. Feinman R, Curtin RR, Shintre S, Gardner AB. Detecting adversarial samples from artifacts. *arXiv preprint* 2017;arXiv:1703.00410
49. Lu J, Issaranon T, Forsyth D. *SafetyNet: detecting and rejecting adversarial examples robustly.* In Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE 2017:446-454
50. Strauss T, Hanselmann M, Junginger A, Ulmer H. Ensemble methods as a defense to adversarial perturbations against deep neural networks. *arXiv preprint* 2017;arXiv:1709.03423
51. Buckman J, Roy A, Raffel C, Goodfellow I. *Thermometer encoding: one hot way to resist adversarial examples.* In ICLR 2018 Conference. La Jolla: ICLR 2018
52. Zantedeschi V, Nicolae MI, Rawat A. Efficient defenses against adversarial attacks. *arXiv preprint* 2017;arXiv:1707.06728
53. Sun S, Yeh CF, Ostendorf M, Hwang MY, Xie L. Training augmentation with adversarial examples for robust speech recognition. *arXiv preprint* 2018;arXiv:1806.02782
54. Su J, Vargas DV, Kouichi S. One pixel attack for fooling deep neural networks. *arXiv preprint* 2012;arXiv:1710.08864
55. Sarkar S, Bansal A, Mahbub U, Chellappa R. UPSET and ANGRI : breaking high performance image classifiers. *arXiv preprint* 2017;arXiv:1707.01159
56. Papernot N, McDaniel P, Goodfellow I, Jha S, Berkay Celik Z, Swami A. Practical black-box attacks against machine learning. *arXiv preprint* 2016;arXiv:1602.02697

딥러닝 기반 의료 영상 인공지능 모델의 취약성: 적대적 공격

김휘영[1] · 정대철[1,2] · 최병욱[1,2*]

딥러닝 학습모델 성능의 비약적인 발전으로 인해 영상의학을 중심으로 하여 기계학습 모델들이 실제 임상현장에서 의사를 보조하여 진단능을 높이고 작업의 효율성을 증대 시켜줄 것이라는 기대가 많다. 이러한 기대로 인해 많은 병원과 민간 기업 등에서 의학 영상을 이용한 자동 진단 학습모델 개발 경쟁이 뜨겁다. 실제로 가까운 미래에 많은 딥러닝 기반의 자동 진단 프로그램들이 의료 환경에서 사용될 것이다. 그러나, 딥러닝 알고리듬이 내재적으로 가진 불확실성(uncertainty)에 의한 적대적 공격(adversarial attack)의 가능성은 특히 의학문제에 딥러닝 알고리듬을 적용하는 데에 큰 걸림돌이 된다. 본 종설에서는 의학영상을 다루는 딥러닝 모델들에 대해 어떠한 원리와 방식으로 적대적 공격이 이루어질 수 있으며, 이로 인하여 어떤 문제들이 발생할 수 있으며 적대적 공격을 차단할 수 있는 방법은 없는지 자세히 살펴보고자 한다.

[1]연세대학교 의과대학 의료영상데이터사이언스센터, 방사선의과학연구소, 영상의학과,
[2]연세대학교 세브란스병원 영상의학과