의료 영상에 대한 적대적 공격 기술의 연구

박윤지 [1], 민지수 [2], 박린 [3], 양진 [4] ,이규영 [5]
[1] 단국대학교 소프트웨어학과 학부생
[2] 가톨릭대학교 인공지능학과 학부생
[3] 수원대학교 정보보호학과 학부생
[4] 서울여자대학교 정보보호학과 학부생
[5] 한국과학기술원 정보보호대학원 박사수료

pyjione@naver.com, zoo5559203@gmail.com, rynpark07@gmail.com, yjswu@swu.ac.kr,leeahn1223@kaist.ac.kr

A Study on Adversarial Attack Techniques in Medical Imaging

Yun-Ji Park[1], Ji-Su Min[2], Ryn Park[3], Jin Yang[4] ,Gyu-young Lee[5]
[1]Dept. of Software, Dan-Kook University
[2]Dept. of Artificial Intelligence, The Catholic University of Korea
[3]Dept. of Information Security, Suwon University
[4]Dept. of Information Security, Seoul Women's University
[5]Graduate School of Information Security, KAIST

요 약

적대적 공격은 영상에 사람의 육안으로는 구분할 수 없는 섭동을 넣어 딥러닝 모델의 오판을 유도한다. 이러한 기술은 의료 분야에서 보험사기 등으로 악용될 수 있고, 이를 해결하기 위해서는 적대적 공격의 원리와 방어기법을 상시 연구하는 노력이 필요하다. 본 논문에서는 다양한 적대적 공격을 코드로 구현하고, 이를 활용하여 의료 CT 영상에 공격을 가하는 실험을 수행하였으며, 그 결과 공격유형 별 원리와 특장점 및 성능 등을 종합적으로 평가하여 방어연구를 촉진하였다.

1. 서론

딥러닝의 발전으로 의료 분야에서 컴퓨터 비전 기술이 전문의 진단을 보조하는 데 널리 활용되고 있다. 그러나 딥러닝은 적대적 공격에 매우 취약하기 때문에, AI 기반 의료시스템은 이에 대한 연구가 필수적이다[1]. 본 논문에서는 가상의 CT 영상 판독 시스템을 대상으로 적대적 공격의 위험성을 평가하고 방어 연구를 촉진하고자 한다.

이를 위해 FGSM, JSMA, DeepFool, C&W 등 대표적인 적대적 공격기술들을 의료 CT 영상에 가하는 비교실험을 수행하였고, C&W 와 JSMA 기법이 가장 높은 성공률을 보였다. 이러한 결과를 바탕으로 각 공격 기법의 원리를 분석하고, 맞춤형 방어전략을 개발함으로써 CT 영상에 대한 적대적 공격을 효과적으로 차단할 수 있으며, 의료 보험사기와 같은 범죄의 예방에도 기여할 수 있을 것으로 예상한다.

2. 관련 연구

본 연구에서는 대표적인 적대적 공격 기법 중 효율성과 정확도를 고려하여 FGSM, JSMA, DeepFool, C&W 공격 기법을 선정하고 실험하였다. 각 공격 기법에 대한 이론적 핵심내용은 다음과 같다.

2.1 FGSM (Fast Gradient Sign Method)

FGSM 공격은 가장 기본적인 적대적 예제 생성 기법이며 단 한 번의 연산만으로 적대적 예제를 생성한다.

$$\boldsymbol{adv_x = x + \varepsilon * sign(\nabla_x J(\theta, x, y))} \quad \textbf{(1)}$$

식(1)에서 adv_x는 생성된 적대적 예제, x는 입력 영상, y는 실측 레이블, ε는 섭동, θ는 신경망 모델, J는 손실 함수를 의미한다. 분류 모델의 손실 함수에 대해 입력영상(x)의 gradient(기울기)를 계산한 후, 그 부호 방향으로 ϵ 만큼의 일정한 섭동을 적용하는 방식이다.

2.2 JSMA (Jacobian Saliency Map Attack)

$$S(\boldsymbol{X}, t)[i] = \begin{cases} 0 \;\; if \;\; \frac{\partial \boldsymbol{F_t(X)}}{\partial \boldsymbol{X_i}} < 0 \;\; or \sum_{j \neq t} \frac{\partial \boldsymbol{F_j(X)}}{\partial \boldsymbol{X_i}} > 0 \\ \left(\frac{\partial \boldsymbol{F_t(X)}}{\partial \boldsymbol{X_i}}\right) \left| \sum_{j \neq t} \frac{\partial \boldsymbol{F_j(X)}}{\partial \boldsymbol{X_i}} \right| \;\; otherwise \end{cases} \quad \textbf{(2) [2]}$$

JSMA 공격은 Jacobian matrix 와 saliency map 을 활용하여 목표 클래스의 확률을 극대화하고 다른 클래스의 확률을 최소화하는 픽셀 쌍을 선택하여 saliency 값을 극대화하는 데 중점을 두는 목표 지정 공격이다.

2.3 DeepFool

DeepFool 공격은 분류 모델이 결정 경계를 넘을 수 있는 최소한의 교란을 추가하여 적대적 예제를 생성한다.

$$\Delta(x;\hat{k}) := \min_{r} |r|_2 \text{ subject to } \hat{k}(x+r) \neq \hat{k}(x) \qquad (3)$$

식(3)에서, 선형 분류기에서는 모델의 출력인 f(x)가 0인 지점이 결정 경계이다. 교란 벡터 r은 이 경계를 넘기 위해 모델의 가중치 벡터 방향을 기반으로 계산된다[3].

2.4 C&W (Carlini & Wagner)

C&W 공격은 방어 기법을 우회할 만큼 강력하며, 계산 과정이 복잡하지만 성공률이 매우 높다.

$$\text{minimize } \mathbf{D}(x, x+\delta) + \mathbf{c}\cdot\mathbf{f}(x+\delta) \qquad (4)$$

식(4)에서, 이 공격은 원본과 적대적 예제의 벡터 간의 거리(D)를 최소화하여 원본과 적대적 예제를 시각적으로 유사하게 만들고, 목적 함수 f를 통해 적대적 예제가 타겟 클래스로 분류되게 만드는 변형(δ)을 찾는다.

$$f_6(x') = \left(\max_{i\neq t}(Z(x')_i) - Z(x')_t\right)^+ \qquad (5)$$

식(5)에서, 목적 함수 $f_6(x')$는 타겟 클래스의 logit 값이 다른 모든 클래스들보다 높아지도록 조정해 적대적 예제가 타겟 클래스로 잘못 분류되도록 유도한다[4].

3. 제안 모델

의료 CT 영상을 사전 학습한 EfficientNetB3, ResNet50, InceptionV3 모델들을 앙상블하여 얻은 모델에 총 100장의 CT영상을 투입하여 실험을 진행한다.

CT 영상 100장에 적대적 공격을 가한 후, 공격 성공률, 차분 영상을 기반으로 측정한 평균 변화 픽셀 수, 적대적 영상을 1장 생성하는데 소요되는 평균 소요시간을 비교하여, 의료 영상에 가장 위력이 높은 공격기법을 찾아낸다.

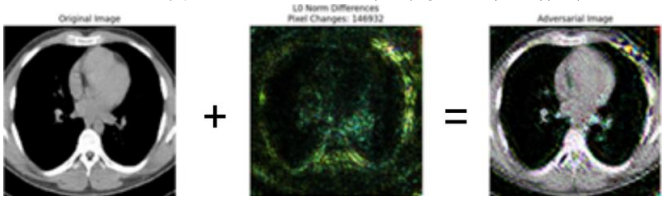
4. 실험 및 고찰

표1과 같이, FGSM은 공격성공률이 64%로 상당히 낮고 투입한 섭동량(1.5197)과 변경한 픽셀의 수(2604.37)가 무척 많았으나, 생성시간은 불과 8.17초로 매우 짧았다.

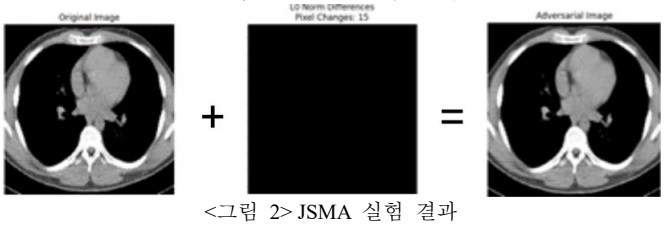
<표 1> 각 공격 기법 별 성능 비교 (평균치)

공격기법	공격 성공률	투입 섭동량	변경 픽셀수	생성시간(초)
FGSM	64%	1.5197	2604.37	8.17
JSMA	90%	0.3820	2.94	75.51
DeepFool	80%	2.1503	13710.15	14.49
C&W	90%	0.0522	17.95	192.33

JSMA와 C&W은 90%의 높은 성공률을 보였고, 투입섭동량은 0.382 및 0.0522 그리고 변경한 픽셀수는 2.94 및 17.95로 FGSM에 비해 수백분의 1에 불과하였으나, 생성시간은 75.51초 및 192.33초로 9배 이상 오래 걸렸다.

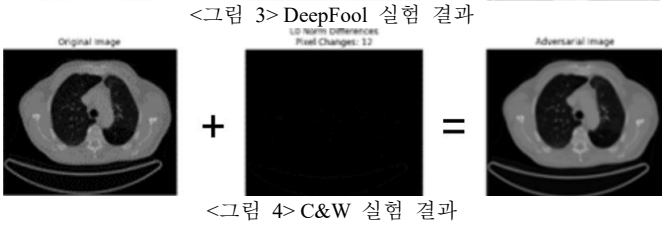


<그림 1> FGSM 실험 결과



<그림 2> JSMA 실험 결과

<그림 3> DeepFool 실험 결과



<그림 4> C&W 실험 결과

그림1부터 4까지의 실험결과에서, 좌측은 원본 영상, 중앙은 원본 영상에 추가한 섭동 차분(적대적영상-원본영상) 영상, 우측은 생성된 적대적 영상을 나타낸다.

FGSM의 섭동차분영상은 사람의 육안으로도 충분히 식별할 수 있을 정도의 많은 섭동이 추가된 것을 볼 수 있으나, 다른 공격기법들은 섭동차분영상이 깨끗하고 육안으로는 원본영상과의 차이를 거의 구별하기 불가할 정도로 적대적 영상이 정교하게 생성된 것을 볼 수 있다.

5. 결론

본 논문에서는 의료 CT 영상에 대하여, 다양한 적대적 공격영상 생성기법을 구현하고 그 성능을 비교하였다.

결론적으로, 생성속도 면에서는 FGSM이 우수하였으나, 변경을 가한 픽셀의 개수와 공격성공율의 측면에서는 JSMA와 C&W 이 훨씬 우수하였다. 따라서 JSMA와 C&W 공격 기법의 잠재된 위험성이 매우 크고 이에 특화된 방어기법 연구가 시급하다는 것을 알 수 있다.

향후 연구에서는 GAN을 활용해 적대적 예제를 Denoise하는 Defense-GAN과 소프트 라벨을 통해 모델의 gradient를 줄여 공격을 무력화하고 모델의 일반화 능력을 향상시키는 Defensive Distillation과 같은 방어 기법을 연구하여 적대적 공격영상에 대한 위험성을 감소시키는 방안을 모색할 예정이다.

ACKNOWLEDGEMENT

※ 본 논문에 있는 학부생들은 모두 공동1저자이며, 논문 작성에 기여한 정도가 같습니다.

※ 본 논문은 과학기술정보통신부 대학디지털교육역량강화 사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] 김휘영,"딥러닝 기반 의료 영상 인공지능 모델의 취약성: 적대적 공격," 대한영상의학회지, vol. 80, 2019.

[2] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," IEEE, 2016.

[3] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard "DeepFool: a simple and accurate method to fool deep neural networks" arXiv, 2015

[4] Nicholas Carlini and David Wagner, "Towards Evaluating the Robustness of Neural Networks," University of California, Berkeley, 2017.