의료 AI 모델 보안성을 위한 적대적 공격 기법 분석 및 최적화 연구

이서윤[1], 김미지[2], 김선아[3], 문소현[4], 정재연[5], 안현주[6]
[1]서울여자대학교 데이터사이언스학과 학부생
[2]숙명여자대학교 통계학과 학부생
[3]충북대학교 미생물학과, 소프트웨어학과 학부생
[4]동국대학교 AI융합학부 학부생
[5]세종대학교 정보보호학과 학부생
[6]리안기술사사무소
winterlike13@swu.ac.kr, kmj25b@sookmyung.ac.kr, seon2134@chungbuk.ac.kr,
munso03@dgu.ac.kr, wodus602@gmail.com, suzic@nate.com

Analysis and Optimization of Adversarial Attack Techniques for the Security of Medical AI Models

Seo-Yun Yi[1], Mi-Ji Kim[2], Seon-Ah Kim[3], So-Hyun Mun[4], Jae-Yeon Jeong[5], Hyun-Joo An[6]
[1]Dept. of Data Science, Seoul Women's University
[2]Dept. of Statistics, Sookmyung Women's University
[3]Dept. of Microbiology, Software, Chungbuk National University
[4]Dept. of AI Convergence, Dongguk University
[5]Dept. of Computer and Information Security·Software, Sejong University
[6]LeeAhn Professional Engineer's Office

요 약

적대적 공격은 의료 영상 모델의 예측을 교란하여 보안성과 신뢰성에 중대한 위협을 가할 수 있다. 본 연구는 의료 영상을 대상으로 네 가지 대표적 적대적 공격 기법을 구현·비교하고, 이 중 가장 강력한 기법에 대해 픽셀 수준의 변경 효과를 실험적으로 분석한다. 이러한 접근은 의료 AI의 보안 취약성 이해를 심화하고, 향후 방어 기법 개발과 임상 적용 가능성 제고에 기여할 것으로 기대된다.

1. 서론

딥러닝 기반 기법은 MRI 등 의료 영상에서 우수한 분류·검출 성능을 보여 임상 보조 도구로서의 활용 가능성이 증가하고 있다. 반면, 최근의 연구들은 적대적 교란(adversarial perturbation)이 매우 작은 변화만으로도 모델의 예측을 크게 변형시킬 수 있음을 보여주어, 의료 영상 시스템의 안정성 및 신뢰성에 대한 우려를 제기하고 있다[1]. 기존 연구들이 주로 전반적인 취약성의 존재 여부를 입증하는 데 초점을 맞췄다면, 본 연구는 보다 구체적으로 서로 다른 공격 기법의 실용적 특성(공격 성공률·연산 시간 등)과, 관심 영역(ROI) 내에서의 픽셀 단위 변경 규모가 실제 분류 성능에 미치는 영향을 정량적으로 평가하는 데 목적을 둔다. 이를 위해 뇌종양 MRI 데이터셋을 사용하여 FGSM, JSMA, Square Attack, ZOO attack을 동일한 실험 환경에서 구현·비교하고, 공격 성능이 우수한 기법에 대해 픽셀 변경 수를 변수로 하는 민감도 실험을 수행한다. 본 연구의 결과는 각 공격 기법의 응용 가능성과 실무적 검증 설계를 위한 기초 자료로 활용될 수 있다.

2. 관련 연구

2.1 FGSM(Fast Gradient Sign Method)

FGSM은 한 번의 계산으로 적대적 이미지를 생성하는 공격 기법으로서, 식(1)과 같이 분류 모델의 손실 함수에 대해 입력영상의 그래디언트(기울기)를 계산한 후, 그 부호 방향으로 ϵ만큼의 일정한 섭동을 적용하는 방식이다.

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x J(\theta, x, y)) \quad (1)$$

2.2 JSMA(Jacobian Saliency Map Attack)

JSMA는 입력 이미지의 각 픽셀이 출력 확률에 미치는 민감도를 계산한 뒤, 그 민감도에 근거해 공격에 가장 효과적인 픽셀 쌍을 차례로 변경하는 표적 공격이다. 민감도 계산은 네트워크 출력의 각 성분에 대한 입력의 편미분들을 모아 만든 자코비안 행렬(Jacobian matrix)에 기반한다. Saliency map은 식(2)와 같이 특정 목표 클래스의 확률을 증가시키는 성분과 다른 클래스들의 합이 갖는 부호 및 크기 정보를 결합해 정의되며, 이 값이 큰 픽셀(또는 픽셀 쌍)이 공격 후보로 선택된다. 선택된 픽셀들에 대해 미리 정한 증분만큼 값을 조정한 뒤, 공격 성공 또는 사전 정의된 예산인 최대 변경 횟수 및 허용 노이즈 한계에 도달할 때까지 이 절차를 반복한다.

$$S(x,t)[i] = \begin{cases} 0, & \text{if } \frac{\partial F_i(x)}{\partial x_i} < 0 \quad \text{or} \quad \sum_{j \neq t} \frac{\partial F_j(x)}{\partial x_i} > 0 \\ \frac{\partial F_i(x)}{\partial x_i} \cdot \left| \sum_{j \neq t} \frac{\partial F_j(x)}{\partial x_i} \right|, & \text{otherwise} \end{cases} \quad (2)[2]$$

2.3 Square Attack

Square Attack은 기울기 정보를 사용하지 않는 기법이다. 해당 기법은 입력 이미지에 각 반복 횟수마다 무작위로 정사각형 패치를 교란하여 적대적 이미지를 생성한다.

3. 제안 기법

3.1 공격기법에 대한 평가지표

본 연구는 MRI 영상 이미지의 이진 분류 환경에서 적대적 공격 기법의 위험성을 정량적으로 평가하고자 공격 성공률, 이미지 당 평균 변경된 픽셀 수, 생성 소요 시간을 지표로 설정하여 종합적으로 분석한다.

3.2 새로운 평가지표의 제안

또한, 본 연구는 모델 예측 신뢰도의 감소 폭을 정량화하기 위해 Confidence Margin Drop(CMD) 지표를 제안한다. 기존 연구에서는 정답-비정답 클래스 간의 로짓 마진을 활용한 목적함수를 통해 적대적 예시를 생성하는 방식이 제안되었다[3]. 이를 확장하여, 본 연구는 이진 분류 문제에서의 CMD를 공격 전후 최대 로짓 값의 차이로 정의한다. 식(3)의 $z(x)$는 입력 x에 대한 모델의 로짓 벡터를 의미한다.

$$CMD = \max(z(x)) - \max(z(x^{adv})) \quad (3)$$

CMD는 단순한 클래스 변경 여부에 그치지 않고, 모델의 확신 저하 정도를 수치화하여 보안 취약성에 대한 평가를 강화하는 지표로 활용될 수 있다.

4. 실험

4.1 실험 환경

Kaggle 데이터 Brain Tumor MRI Dataset[4]을 기반으로, 사전 학습한 ResNet-50 모델에 대해 랜덤 샘플링한 100장의 테스트 이미지를 선정하였다. 이를 대상으로 FGSM, JSMA, Square Attack, ZOO Attack 4가지 적대적 공격을 수행한다.

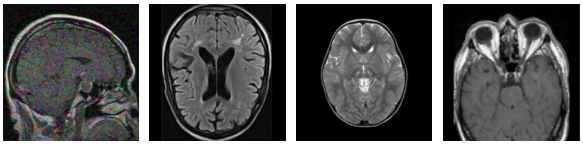
4.2 실험 결과 및 고찰

<표 1>과 같이 FGSM(69.00%)과 ZOO Attack(67.00%)에 비해 JSMA(89.00%)와 Square Attack(88.00%)이 상대적으로 높은 성공률을 보였다.

<표 1> 각 공격기법 별 성능 (평균)

공격기법	성공률	총 변경 픽셀 수	시간(초)	CMD
FGSM	69.00%	179,098.21	0.02	15.99
JSMA	89.00%	1,415.91	45.55	14.20
Square	88.00%	23,557.63	19.54	15.83
ZOO	67.00%	128,500.00	83.89	12.94

FGSM은 변경 픽셀 수가 많았으나(179,098.21개) 적대적 이미지 생성 시간은 매우 짧았다(0.02초). 반면, JSMA는 변경 픽셀 수가 가장 적었으나(1,415.91개), 생성 시간은 45.55초로 비교적 긴 시간이 소요되었다. 4가지 기법 모두 CMD 지표를 통해 모델의 신뢰도가 감소한 것을 확인하였다.



(그림2) FGSM, JSMA, Square, ZOO 공격 후 이미지

가장 높은 공격 성공률과 적은 픽셀 변경으로 효율성을 보인 JSMA 기법에 대해, 소요 시간 단축을 위한 픽셀 수(k) 변경 실험을 수행하였다. JSMA의 기본 설정인 k=2 대비, k=5, 10으로 확장한 결과, <표 2>와 같이 공격 성공률은 큰 변화가 없었으나, 소요 시간은 약 9배 단축되었다.

<표 2> 각 변경 픽셀 수 별 JSMA 성능 (평균)

픽셀 수	성공률	총 변경 픽셀 수	시간(초)	CMD
k=2	89.00%	1415.91	45.55	14.20
k=5	88.00%	1439.39	9.13	14.33
k=10	87.00%	1449.54	5.00	14.21

5. 결론 및 고찰

본 연구는 의료 CT 및 뇌종양 MRI 영상 이미지를 대상으로 FGSM, JSMA, Square Attack, ZOO의 성능을 비교하고 분석하였다. 결과적으로 FGSM은 생성 속도가 빠른 반면, 동일한 교란 예산(변경 픽셀 수)에서는 JSMA가 더 높은 공격 성공률을 보였다. 또한 JSMA의 회당 변경 픽셀 수를 늘리는 간단한 운용 최적화가 공격 성공률을 유지하면서 전체 생성 시간을 크게 단축함을 확인하였다. Square Attack은 블랙박스 환경에서도 실용적 위협이 될 가능성이 있어, JSMA에 대한 특화된 방어와 함께 Square Attack의 쿼리 효율·방어 회피 특성에 대한 추가 평가가 필요하다.

향후에는 ROI 기반 표적화, Defense-GAN·Defensive Distillation 및 Adversarial training·Certified defenses와 같은 다층적 방어 전략을 통해 의료 영상 시스템의 실무적 강인성을 확보하는 방향으로 연구를 확장할 예정이다.

ACKNOWLEDGEMENT

※ 본 논문은 과학기술정보통신부 대학디지털교육역량강화사업이 지원한 한이음 드림업 프로젝트의 결과물입니다.

※ 본 논문에 있는 학부생들은 모두 공동 1저자이며, 논문 작성에 기여한 정도가 같습니다.

참고문헌

[1] Tsai MJ, Lin PY, Lee ME, “Adversarial Attacks on Medical Image Classification.” Cancers, 15, 17, 2023,

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

[3] N. Carlini and D. Wagner, “Towards Evaluating the Robustness of Neural Networks,” 2017 IEEE Symposium on Security and Privacy (SP), 2017.

[4] Msoud Nickparvar, “Brain Tumor MRI Dataset [Data set],” Kaggle, 2021.