Adaptive ε 스케줄링 기반 PGD 전처리를 이용한 딥페이크 방어 기법

양유진*, 김지태, 강민서, 오권욱, 김동재, 이재욱
국립부경대학교 정보통신공학과
{eu9309*, kjt000110, minseo3734, 202230446}@pukyong.ac.kr, {kdj6306, jlee0315}@pknu.ac.kr

Deepfake Defense Method Using PGD Preprocessing Based on Adaptive ε-Scheduling

Yujin Yang*, Jitae Kim, Minseo Kang, Gwonuk O, Dongjae Kim and Jaewook Lee
Dept Information and Communications Eng., Pukyong National University

요 약

최근 허위 이미지·영상물을 이용한 범죄가 급증하며 디지털 보안 위협이 심화되고 있다. 생성형 AI 모델은 이미지·영상 합성 등 다양한 분야에 혁신을 가져왔지만, 악의적 활용 가능성 또한 증가시켰다. 이에 본 연구에서는 딥페이크 방어를 위해 ResNet-18 기반 분류기를 제작·튜닝하고, PGD(Projected Gradient Descent) 공격과 적응형(Adaptive) ε 스케줄링 기법을 결합한 이미지 사전 처리 알고리즘을 제안한다. 제안 기법은 SSIM(Structural Similarity Index)를 0.95 이상으로 유지하면서, 최소한의 ε 값으로도 높은 공격 성공률을 달성하도록 설계되었다. 다양한 ε 실험을 통해 기존 단일-스텝 FGSM(Fast Gradient Sign Method) 대비 성공률 및 이미지 품질 유지 측면에서 우수함을 검증하였다.

Ⅰ. 서 론

최근 이미지 처리 분야에서는 GAN 및 CNN을 기반으로 한 생성형 AI 모델이 컴퓨터 그래픽(CG), 엔터테인먼트, 광고 등 다양한 산업에서 혁신적인 영상·이미지 제작 도구로 자리매김해왔다. 그러나 이러한 기술이 악의적으로 활용될 경우, 사회적·윤리적 문제를 야기할 수 있다. 딥페이크를 이용한 허위 영상 범죄는 2021년 대비 약 197% 증가하여 그 심각성이 크게 대두되고 있다. 이러한 허위 영상물 문제 해결을 위해 FWA 기반의 왜곡 분석 기법, 심층 신경망 기반 탐지 모델(EfficientNet 등)을 활용한 포렌식 방법 등이 제안되었다. 그러나 이들 대부분은 사후 탐지 및 차단을 목표로 하므로, 사용자가 이미지를 업로드하기 전에 악용을 차단할 수 있는 사전 예방적 대응 기술의 필요성이 강조된다.[1]

본 연구에서는 PGD와 적응형 ε 스케줄링 기법을 결합한 이미지 사전 처리 알고리즘을 제안한다. 기존의 단일-스텝 FGSM 사전 처리를 통한 딥페이크 방어 기법은 간단하고 빠르지만, 픽셀마다 불연속 노이즈가 발생하여 이미지 품질(SSIM)이 크게 저하되는 한계가 있다. 제안 기법은 미세 스텝을 수 회 반복하여 overshoot를 완화하고 공격 성공과 SSIM ≥ 0.95 두 가지 조건을 동시에 만족하는 최소 ε를 자동 탐색하는 Adaptive ε 스케줄링을 통해 최소한의 perturbation으로도 높은 공격 성공률을 달성한다.

사용자는 제안 기법을 통해 화질 저하를 최소화하면서 사전 처리된 이미지를 이용하여 딥페이크 모델에 혼란을 주어 SNS을 통한 딥페이크 피해를 방어할 수 있다.[1] 또한, 본 알고리즘을 웹 플랫폼 또는 SNS 업로드 파이프라인에 적용할 경우 사용자가 직접 복잡한 설정 없이 노이즈를 삽입하여 AI를 이용한 합성을 방지할 수 있으므로 향후 디지털 콘텐츠의 보안성을 크게 향상시킬 수 있을 것으로 기대된다.

Ⅱ. 알고리즘 제안

본 절에서는 제안하는 알고리즘의 구조와 설계 과정을 설명한다. 그림 1은 원본 이미지에 대해 PGD로 미세 스텝을 반복 적용한 뒤 각 스텝에서 projection 및 클램핑을 수행하는 전체적인 흐름을 나타낸다.

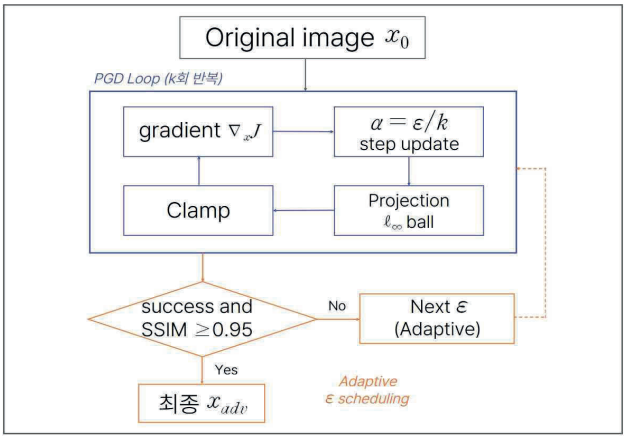

그림 1. 알고리즘 구조

2-1. Multi-step PGD

PGD는 단일-스텝 FGSM 한계를 극복하기 위해 ε을 한 번에 모두 적용하지 않고, ε/k 크기의 미세 스텝을 k회 반복 수행하여 적대적 예제를 생성하는 기법이다. 각 반복에서 손실 함수 $J(x_{adv}, y)$를 x_{adv}에 대해 미분해 gradient $\nabla_x J$를 계산하는 스텝 업데이트 과정은 다음과 같이 정의된다.

$$x_{adv} \leftarrow x_{adv} + \alpha \cdot sign(\nabla_x J),\ \alpha = \frac{\varepsilon}{k}$$

원본 이미지 x_0와의 $\ell\infty$-거리 $\| x_{adv} - x_0 \|_\infty \le \varepsilon$를 유지하도록 하는 식은 다음과 같이 정의된다.

$$x_{adv} \leftarrow clip(x_{adv}, x_0 - \varepsilon, x_0 + \varepsilon)$$

각 채널별 정규화 범위 $[(0-\mu)/\sigma, (1-\mu)/\sigma]$로 다시 클램핑하여, 모델 입력이 유효한 픽셀 값으로 유지되도록 한다. 이 과정을 통해 한 번에 과도하게 업데이트 되는 overshoot 현상을 완화하고, ε 증가에 따라 단조하게 공격 성공률이 증가하도록 한다.

2-2. Adaptive ε 스케줄링

Adaptive ε 스케줄링은 여러 후보 ε 값을 순차적으로 시험해 보고, "적대적 공격 성공"과 "SSIM ≥ 0.95" 두 조건을 동시에 만족하는 최소 ε를 자동으로 선택하는 기법이다.[2]

먼저 $\{\varepsilon_1 < \varepsilon_2 < \cdots < \varepsilon_n\}$ 형태로, 0.005부터 0.1까지 ε 리스트를 생성한다. 각 ε에 대해 PGD를 수행하여 $x_{adc}(\varepsilon)$를 얻고, 모델의 결과 예측이 변경되는지 확인한다. 또한, $SSIM(x_0, x_{adv}) \geq 0.95$인지 계산한다. 두 조건을 만족하는 첫 번째 ε를 선택하며, 이 ε는 최소한의 perturbation으로도 높은 성공률을 보장하고 SSIM 손실을 최소화한다. Adaptive ε 스케줄링을 통해, 사용자는 일률적인 ε 설정 없이도 이미지 품질(SSIM)과 공격 효율 간 최적의 trade-off를 자동으로 달성할 수 있다.

Ⅲ. 실험 결과

본 실험에서는 직접 제작한 custom 데이터셋(600장)을 대상으로, 단일-스텝 FGSM과 제안한 PGD+Adaptive ε 스케줄링 기법의 공격 성공률을 비교하였다.

3-1. 공격 성공률 분석

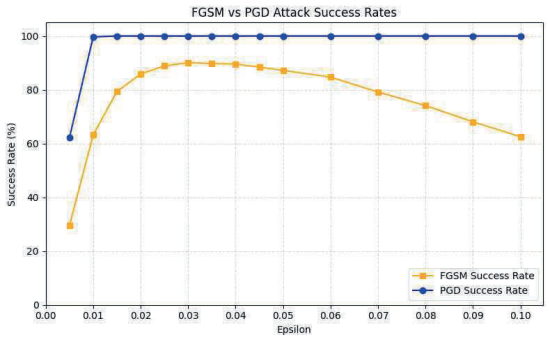

그림 2. 공격 성공률 비교

그림 2는 ε값을 0.005에서 0.1까지 증가시키며 단일-스텝 FGSM(파란 실선)과 제안한 PGD+Adaptive ε 스케줄링(빨간 실선) 기법의 공격 성공률을 나란히 비교한 결과이다. FGSM은 $\varepsilon \approx 0.03$ 부근에서 90.08%까지 오른 뒤 0.07 지점에서 다시 79.33%까지 떨어지는 비단조적(dip) 패턴을 보이나, 제안 기법은 ε=0.007에서 이미 90.32%의 높은 성공률을 달성하고 ε 증가에 따라 단조적으로 상승하여 ε=0.01에서 99.66%에 도달한다.

3-2. SSIM 보존 성능

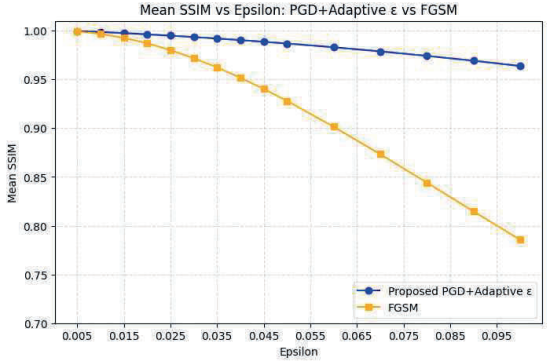

그림 3. 평균 SSIM 비교

평균 SSIM 측정 결과, ε=0.07일 때, FGSM은 SSIM=0.84까지 떨어진 반면, 제안 기법은 SSIM=0.96을 유지하였다. 또한, 제안 기법은 그래프 내 모든 범위에서 SSIM ≥ 0.95 조건이 안정적으로 유지되었다.

아래 예제에서는 동일 원본 이미지를 대상으로, 그림 4. FGSM(ε=0.03, SSIM=0.8755)과 그림 5. 제안 PGD+Adaptive ε(ε=0.007, SSIM=0.9996) 이미지 생성 결과를 비교하였다. 제안 기법은 고품질을 유지하면서도 성공적인 공격을 달성하였다.

그림 4.
기존 FGSM

그림 5.
PGD+adaptive ε

3-3. ε 선택 분포

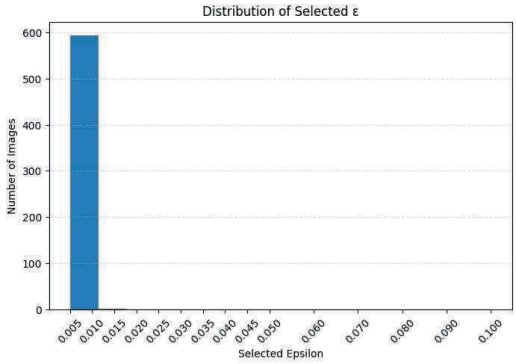

그림 6. selected ε

Adaptive ε 스케줄링으로 600장의 이미지마다 선택된 ε를 집계한 결과, 전체의 85%가 $\varepsilon \in [0.005, 0.012]$ 구간에서 최소 ε가 결정되었고, 15%는 $\varepsilon \in [0.012, 0.017]$ 구간이었다. 그림 6에서 ε가 매우 작은 값에서 집중되는 모습을 보여, 실제 적용 시 대부분의 이미지에 대해 매우 작은 perturbation으로도 공격 성공과 SSIM 제약을 동시에 만족함을 보여준다.

Ⅳ. 결 론

본 연구에서는 허위 이미지·영상의 악용을 사전 차단하기 위해, PGD와 Adaptive ε 스케줄링을 결합한 사전 처리 알고리즘을 제안하였다. ε/k 크기의 미세 스텝을 반복 적용하고, ε 값을 최소화하면서 SSIM ≥ 0.95를 보장하는 ε를 자동 탐색함으로써, 기존 FGSM 대비 적은 ε로도 높은 공격 성공률을 달성하였다.

실험 결과, FGSM은 $\varepsilon \approx 0.03$에서 90.08% 정도의 성공률을 보이며 SSIM이 0.80 이하로 급락했던 반면, 제안 기법은 $\varepsilon \approx 0.007$에서 이미 90% 이상의 성공률을 기록하고 SSIM을 0.9966 이상으로 유지하였다. 또한 ε 분포 분석에서 대다수(≈85%)의 이미지가 $\varepsilon \in [0.005, 0.012]$ 구간에서 충분히 공격이 성공하는 것으로 나타나, 실제 적용 시에도 매우 작은 perturbation으로 실용적 방어 효과를 기대할 수 있음을 확인했다.

참 고 문 헌

[1] Yujin Yang, Jitae Kim, Minseo Kang, Gwonuk O, "Preventing Deepfake Using FGSM Adversarial Attacks," *Korea Institute Of Communication Sciences (KICS)*, Feb. 2025

[2] Liu, Ye; Cheng, Yaya; Gao, Lianli; Liu, Xianglong; Zhang, Qilong; Song, Jingkuan; "Practical Evaluation of Adversarial Robustness via Adaptive Auto Attack," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June. 2022