안티 포렌식에 강인한 딥페이크 탐지 기법

민지민[1], 김지수[1], 김민지[1], 장한얼[1]
[1]한밭대학교 컴퓨터공학과
{20191730, 20191769, 20191767}@edu.hanbat.ac.kr, hejang@hanbat.ac.kr

A Robust Deepfake Detector against Anti-forensics

Ji-Min Min[1], Ji-Soo Kim[1], Min-Ji Kim[1], Haneol Jang[1]
Dept. of Computer Engineering, Hanbat National University

요 약

인공지능 기반의 딥페이크(Deepfakes) 기술이 사회적인 이슈로 대두되고 있다. 하지만 기존 딥페이크 탐지기는 sharpening, additive noise와 같은 간단한 이미지 변형만으로 탐지 우회가 가능한 문제점이 있다. 본 논문에서는 안티 포렌식에 강인한 딥페이크 탐지기를 개발하기 위해 이미지 편집 도구 기반의 안티 포렌식 데이터셋을 생성하고 적대적 학습을 수행하는 방법을 제안한다. 실험 결과를 통해 안티 포렌식에 취약한 기존 딥페이크 탐지기 성능이 제안한 적대적 학습 기법을 수행한 이후에 탐지율이 크게 개선된 것을 확인할 수 있었다.

1. 서론

최근 전 세계적으로 '가짜뉴스' 및 '가짜 연예인 음란 동영상'에 사용되는 인공지능 기반의 딥페이크 기술이 사회적인 이슈로 대두되고 있다. 2019년 네덜란드의 보안업체 Deeptrace의 The state of deepfakes 2019에 따르면 전 세계 온라인 딥페이크 사용의 96%가 포르노그라피라고 한다. 심지어 해당 보고서에 따르면 딥페이크 포르노그라피의 피해자 중 25%가 한국 여성이며 연예인과 일반인 모두 해당한다고 한다. 또한 딥페이크 기술은 빠른 개발 속도와 쉬운 접근성을 기반으로 '개인 정보 침해', '사기' 등 다양한 범죄에 악용되고 있다. 이에 따라 딥페이크 영상 탐지에 대한 연구가 활발히 진행되고 있다.

기존 딥페이크 탐지기는 입력 영상에 sharpening, additive noise와 같이 간단한 이미지 변형을 가하는 공격만으로 무력화되는 문제가 존재한다. 이러한 디지털포렌식 회피 기법을 안티 포렌식(anti-forensics)이라고 한다. 안티 포렌식 기법을 사용하여 딥페이크 탐지기를 우회하는 사례가 등장함에 따라 안티 포렌식 기법에 강인한 딥페이크 탐지 연구가 시급한 상황이다.

본 논문에서는 적대적 학습 기법을 이용한 안티 포렌식에 강인한 딥페이크 탐지 기법을 제안한다.

2. 관련 연구

2.1 안티 포렌식

안티 포렌식이란 디지털포렌식 기술에 대응하여 디지털 데이터를 조작, 삭제, 은닉, 또는 난독화하여 증거를 은폐하는 행위를 말한다.[1]

안티 포렌식 기법 중 데이터 조작 기법을 사용하여 딥러닝의 예측 오류를 최대화하는 기법으로 적대적 공격이 있다. 적대적 공격은 딥러닝의 심층 신경망을 이용한 모델에 교란(perturbation)을 적용하여 궁극적으로 탐지기의 오 분류를 일으키는 공격이다. 그리고 딥러닝은 이러한 적대적 공격에 취약하다고 알려져 있다.[2]

이와 같은 안티 포렌식 데이터셋 생성 기법으로는 크게 화이트박스 공격(White-Box Attacks)과 블랙박스 공격(Black-Box Attacks)이 있다.[3]

2.1.1 화이트박스 공격(White-Box Attacks)

화이트박스 공격은 공격자가 모델에 대한 모든 정보를 알고 공격하는 공격을 의미한다.

화이트박스 공격의 대표적인 알고리즘으로 Fast Gradient Signed Method(FGSM)와 Projected Gradient Descent(PGD)가 있다.

2.1.2 블랙박스 공격(Black-Box Attacks)

블랙박스 공격은 공격자가 모델에 대한 정보를 알지 못하고 공격하는 것을 의미한다. 공격자는 사전에 제작된 값을 입력하여 출력되는 결과를 관찰하며 모델의 취약성을 분석한다.

블랙박스 공격은 비 적응형 블랙박스 공격, 적응형 블랙박스 공격, 엄격한 블랙박스 공격으로 구분된다. [4]

그 중 엄격한 블랙박스 공격을 수행하는 공격자는 입력-출력 데이터를 얻을 수 있지만 출력의 변화를 관찰하기 위해 입력을 변경할 수 없다. 엄격한 블랙박스 공격으로 Universal Perturbation 기법이 있다. Universal Perturbation 기법은 공격자가 모든 원본 데이터에 특정 노이즈를 추가하는 일반적인 공격 기법을 통해 오 분류를 유도하는 방법이다. 최근에는 이미지 편집 도구를 이용한 엄격한 블랙박스 공격 사례가 등장하고 있다.

이미지 편집 도구를 이용한 엄격한 블랙박스 공격 기법의 대표적인 예로 'Gaussian Noise', 'Sharpening', 'JPEG Compression'이 있다.

'Gaussian Noise'는 의도적으로 이미지를 손상시키기 위해 이미지에 가우시안 확률 밀도 함수를 따르는 노이즈 신호를 추가하여 시각적 품질을 떨어뜨리는 이미지 편집 기법이다. [5]

'Sharpening'은 일반적으로 이미지 경계의 대비 효과를 증가시켜 이미지를 선명해지도록 하는 이미지 편집 기법이다. 일반적으로 이미지와 sharpening mask의 컨볼루션 연산을 통해 선명한 출력 이미지를 생성한다. [6], [7]

'JPEG Compression'은 색공간 변환, 다운 샘플링, DCT(Discrete Cosine Transform) 변환, 양자화 및 엔트로피 코딩의 과정을 거쳐 압축된 이미지를 생성하는 이미지 편집 기법이다. 해당 기법에서 조절 가능한 옵션은 다운 샘플링 방법, DCT 방법, 양자화 테이블이 있는데, 이 중 양자화 테이블은 일반적으로 이미지 품질 계수인 Quality Factor(QF)와 대응되며 압축된 출력 이미지의 품질을 조절하기 위해 사용한다. [8]

이미지 편집 도구를 이용한 블랙박스 공격은 비전문가도 쉽게 이용할 수 있기 때문에 위험성이 매우 높은 공격이다.

화이트박스 공격은 모델에 대한 대부분의 정보를 갖고 있기 때문에 공격 성공률이 100%에 가깝지만 모든 정보를 알고 있다는 조건이 비현실적이다. 그로 인해 실제로는 화이트박스 공격 보다는 블랙박스 공격이 많이 시도되고 있다.[9]

2.2 적대적 학습

안티 포렌식 데이터셋 생성과 같은 적대적 공격을 견딜 수 있도록 네트워크를 훈련시키는 기법으로 적대적 학습이 있다. 적대적 학습은 간단한 방법이지만 적대적 사례를 사용하는 공격뿐 만 아니라 블랙박스 공격에 대해서도 방어가 가능하다. 블랙박스 공격에 대한 방어의 경우, 원본 데이터셋에 안티 포렌식 기법을 적용하여 생성한 데이터셋을 원본 데이터셋과 함께 학습하여 노이즈에 상관없이 제대로 인식하도록 한다. 적대적 학습의 목표는 가능한 한 많은 안티 포렌식 기법이 적용된 데이터셋을 모델에 입력하여 적대적 공격에 강인한 모델을 생성하는 것이다.

본 논문에서는 이미지 편집 도구를 이용한 블랙박스 공격에 대해서 적대적 학습을 수행하여 보다 강인한 딥페이크 탐지기를 개발한다.

3. 시스템 제안

이미지 편집 도구를 이용한 안티 포렌식 기법을 원본 데이터셋에 적용하여 안티 포렌식 데이터셋을 생성한다. 생성한 안티 포렌식 데이터셋을 기존 학습 데이터셋에 추가하여 적대적 학습을 수행함으로써 안티 포렌식에 강인한 딥페이크 탐지 모델을 개발한다.

3.1 데이터셋

3.1.1 원본 데이터셋

사람의 얼굴 이미지를 학습 데이터로 사용해야 하므로 공인된 AI Hub의 '딥페이크 변조 영상' 데이터셋을 사용한다.

3.1.2 안티 포렌식 데이터셋 생성

모든 정보를 다 알고 있다는 조건 하에 이루어지는 화이트박스 공격에 비해 블랙박스 공격이 실제 발생 가능성이 높다. 이에 따라 본 논문에서는 안티 포렌식 기법 중 엄격한 블랙박스 공격을 적용하여 안티 포렌식 데이터셋을 생성한다. 적용한 안티 포렌식 데이터셋 생성 기법은 'Gaussian Noise', 'Sharpening', 'JPEG Compression'이다.

발생할 수 있는 다양한 안티 포렌식 공격을 우회하기 위해 **[표 1]**은 'Gaussian Noise', 'Sharpening',

'JPEG Compression(Quality Factor)'를 총 10단계의 강도로 설정하는 파라미터 값을 나타낸다.

Attack Level	Gaussian Noise	Sharpening	JPEG Quality Factor
1	30	3	90
2	60	5	85
3	90	7	75
4	120	9	65
5	150	11	55
6	180	13	45
7	210	15	35
8	240	17	25
9	270	19	15
10	300	21	5

표 1. 안티 포렌식 공격 단계별 파라미터

[그림 1]은 'Gaussian Noise' 기법이 5단계의 강도로 적용된 이미지의 예시이다.

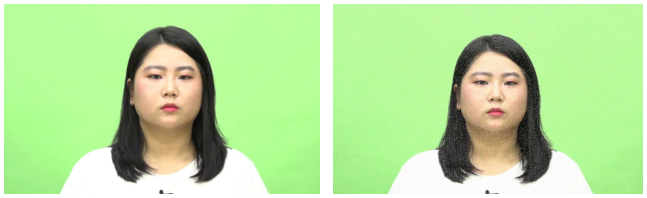

그림 1. Gaussian Noise 적용 예시

[그림 2]는 'Sharpening'이 4단계로 적용된 이미지의 예시이다.

그림 2. Sharpening 적용 예시

[그림 3]는 'JPEG Compression' 기법이 9단계의 강도로 적용된 이미지의 예시이다.

그림 3. JPEG Compression 적용 예시

3.2 모델 학습 기법

본 논문은 **[그림 4]**로 표현할 수 있는 3단계의 모델 학습 기법을 통하여 안티 포렌식에 강인한 모델의 개발을 제안한다.

[a] - 첫 번째 단계에서는 원본 데이터셋만을 사용해 딥페이크 탐지기를 학습한 모델을 도출한다.

[b] - 두 번째 단계에서는 각 안티 포렌식 기법의 파라미터 수치를 조절하여 10단계의 강도로 적용해 안티 포렌식 데이터셋을 생성한다.

[c] - 세 번째 단계에서는 첫 번째 단계에서 도출된 모델의 학습 데이터셋으로 원본 데이터셋과 생성한 안티 포렌식 데이터셋을 함께 포함하는 데이터 증대를 적용한 적대적 학습을 통하여 안티 포렌식에 강인한 모델을 개발한다.

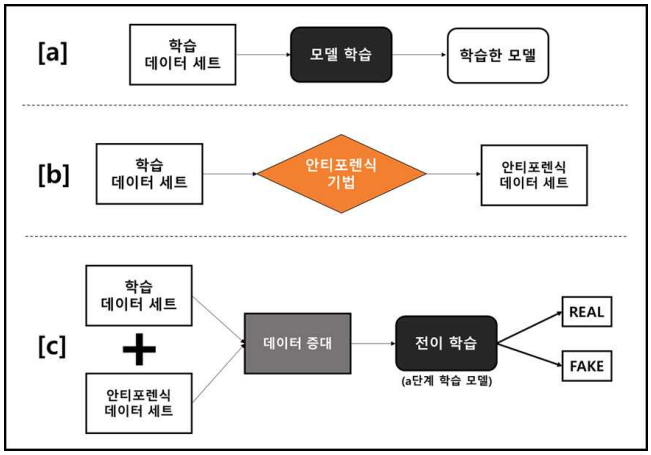

그림 4. 모델 학습 절차

4. 실험 결과 및 결론

원본 데이터셋의 training set은 5000장이고 validation set과 test set은 각 1800장이며, real이미지와 fake이미지의 개수는 동일하다. 안티포렌식 데이터셋의 크기는 원본 데이터셋의 크기와 동일하도록 생성하였다.

모델 학습 시 Epoch은 3으로, 배치 사이즈는 32로 진행하였다. Adam Optimizer를 통해 최적화를 진행하였고, 학습률의 경우 PyTorch의 Learning Rate Scheduler 중 StepLR을 사용하였다.

[표 2]는 Xception[10] 기반 딥페이크 탐지 모델을 이용하여 원본 데이터셋을 학습한 결과이다.

Validation Loss	Validation Accuracy
0.0015	1.000

표 2. 딥페이크 탐지 모델 학습 결과

[그림 5]의 test accuracy 그래프를 통해 안티 포렌식 데이터셋에 대한 탐지 정확도가 현저히 하락했음을 알 수 있고, 이를 통해 학습한 딥페이크 탐지 모델이 안티 포렌식에 취약함을 확인할 수 있었다.

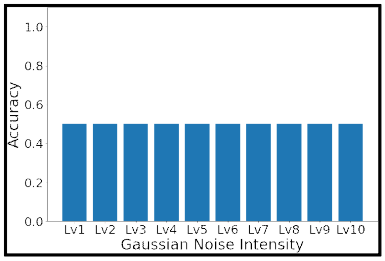

그림 5-1. 'Gaussian Noise' 공격 단계별 딥페이크 탐지 성능

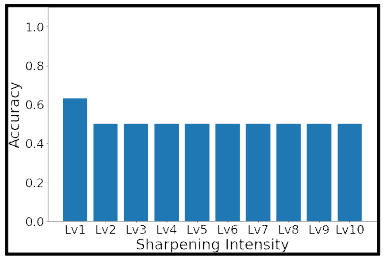

그림 5-2. 'Sharpening' 공격 단계별 딥페이크 탐지 성능

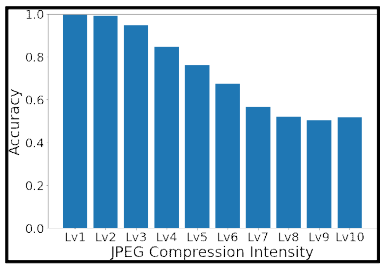

그림 5-3. 'JPEG Compression' 공격 단계별 딥페이크 탐지 성능

[그림 6]의 test accuracy 그래프를 통해 제안한 적대적 학습 기반의 딥페이크 탐지 모델이 안티 포렌식 공격이 가해진 데이터셋에 대해서도 높은 판별 정확도를 보임을 확인할 수 있었다.

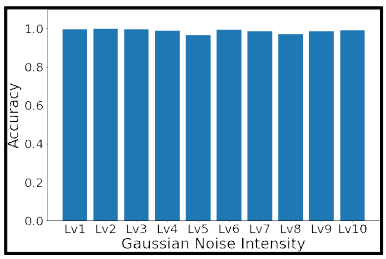

그림 6-1. 'Gaussian Noise' 공격 단계별 적대적 학습 모델 판별 정확도

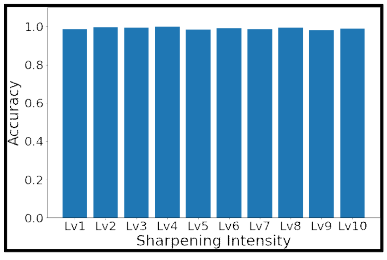

그림 6-2. 'Sharpening' 공격 단계별 적대적 학습 모델 판별 정확도

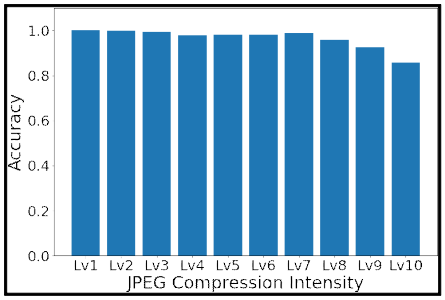

그림 6-3. 'JPEG Compression' 공격 단계별 적대적 학습 모델 판별 정확도

실험 결과를 통해 간단한 이미지 변형에도 취약했던 기존 딥페이크 탐지 모델에 적대적 학습 기법을 적용함으로써 높은 탐지 강인성을 획득할 수 있음을 확인하였다.

향후 학습하지 않은 안티 포렌식 기법에 대해서도 높은 강인성을 갖는 딥페이크 탐지 기법을 연구할 계획이다.

참고문헌

[1] Lee, Seok-Hee, et al. "안티 포렌식 기술과 대응 방향." *Review of KIISC* 18.1 (2008): 11-19.

[2] Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." *Ieee Access* 6 (2018): 14410-14430.

[3] 오유진, et al. "적대적 사례 생성 기법 동향." *한국정보처리학회 학술대회논문집* 28.2 (2021): 580-583.

[4] Chakraborty, Anirban, et al. "Adversarial attacks and defences: A survey." *arXiv preprint arXiv:1810.00069* (2018).

[5] Singh, Prabhishek et al. "Image Watermark Attacks: Classification and Implementation," The International Journal of Electronics & Communication Technology, vol. 4, no. 2, pp. 95-100, 2013.

[5] Jeon, Yong-Tae, Hyun Lee, and Jae-Sung Choi. "Depth Image Based Feature Detection Method Using Hybrid Filter." *IEMEK Journal of Embedded Systems and Applications* 12.6 (2017): 395-403.

[7] Noh, Gyumyung, Seungwoo Wee, and Jechang Jeong. "Image Sharpening Algorithm Using Morphological Operations." *Proceedings of the Korean Society of Broadcast Engineers Conference*. The Korean Institute of Broadcast and Media Engineers, 2019.

[8] 허욱, et al. "JPEG 이미지 압축정보를 이용한 이미지 유포 기기 특정 방안 연구." *디지털포렌식연구* 14.1 (2020): 33-44.

[9] Kwon, Hyun, and Yongchul Kim. "딥러닝 모델에 대한 적대적 사례 기술 동향." Review of KIISC 31.2 (2021): 5-12.

[10] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.