적대적 공격에 따른 딥페이크 탐지 모델 강화

이상영 [1], 허종욱 [1*]
[1] 한림대학교 소프트웨어학부
thf0415@naver.com, juhou@hallym.ac.kr

Improving the Robustness of Deepfake Detection Models Against Adversarial Attacks

Sangyeong Lee[1], Jong-Uk Hou[1*]
[1]Division of Software, Hallym University

요 약

딥페이크(deepfake)로 인한 디지털 범죄는 날로 교묘해지면서 사회적으로 큰 파장을 불러일으키고 있다. 이때, 딥러닝 기반 모델의 오류를 발생시키는 적대적 공격(adversarial attack)의 등장으로 딥페이크를 탐지하는 모델의 취약성이 증가하고 있고, 이는 매우 치명적인 결과를 초래한다. 본 연구에서는 2 가지 방법을 통해 적대적 공격에도 영향을 받지 않는 강인한(robust) 모델을 구축하는 것을 목표로 한다. 모델 강화 기법인 적대적 학습(adversarial training)과 영상처리 기반 방어 기법인 크기 변환(resizing), JPEG 압축을 통해 적대적 공격에 대한 강인성을 입증한다.

1. 서론

소셜 네트워크 서비스에 올렸던 사진이 도용되어 포르노 영상에 합성이 되기도 하고, 유명 인사들의 영상이 조작되어 가짜 뉴스와 같은 결과물이 만들어지기도 한다. 이를 가능하게 한 딥페이크(deepfake) 기술은 계속되어 발전하고 있고, 이제 사람의 눈으로는 딥페이크 여부를 판단하기 힘들어졌다. 딥페이크 기술이 발전됨에 따라 딥페이크를 검출해낼 수 있는 딥러닝 기반 모델 또한 계속해서 연구되고 있다 [1].

하지만, 딥러닝 모델을 무력화시키는 적대적 공격(adversarial attack)의 등장으로 딥러닝 모델은 쉽게 오류를 범할 수 있게 되었다. Szegedy 연구팀[2]에 따르면, 적대적 공격이란 딥러닝을 이용한 모델에 적대적 섭동(adversarial perturbation)을 적용하여 오분류를 발생시키는 것을 말한다. 섭동은 모델의 예측 오류를 최대로 만드는 방향으로 모델 입력값을 최적화하면서 형성되고, 육안으로는 구분하기 어려울 정도로 작고 미세하게 나타난다. 이런 섭동이 더해진 입력 값을 적대적 예제(adversarial example)라고 한다.

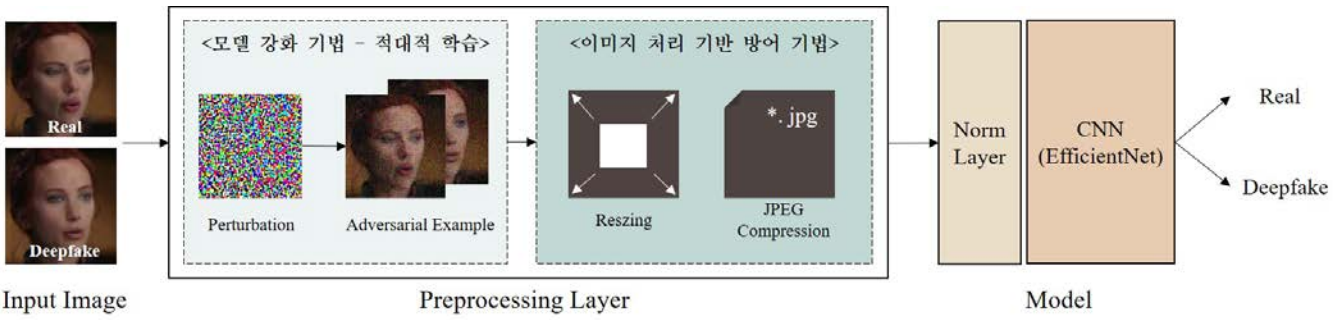
적대적 예제를 생성하는 방법 중 가장 기본적인 방법인 FGSM(Fast Gradient Signed Method)[3]는 신경망의 기울기(gradient)를 이용해 적대적 예제를 생성하는 기법이다. 딥러닝 기반 모델은 학습할 때, 역전파 단계에서 손실 함수를 최소화하도록 모델에 기울기를 적용하여 업데이트한다. FGSM 에서는 이러한 손실 함수를 최대화하는 방향으로 모델이 아닌 입력 이미지에 기울기를 적용하여 모델이 제대로 된 판단을 내릴 수 없게 만든다. 이미 잘 판별하고 있는 모델에 적대적 예제를 입력 값으로 넣을 경우, 모델은 전혀 다른 예측 값을 내보이게 된다. Lin 연구팀[4]은 DNN(Deep Neural Network)에 4 가지의 적대적 공격(FGSM, PGD, BIM, MIM)을 적용하여 딥러닝 기반 모델이 적대적 공격에 매우 취약하다는 것을 증명하였다.

적대적 공격으로 인해 딥페이크 이미지를 일반 이미지로 판단하는 경우 큰 혼란을 일으킬 수 있다. 따라서 적대적 공격에 대한 딥러닝 기반 모델의 견고성은 디지털 포렌식에서 중요한 문제이다. 본 연구에서는 적대적 공격에 대한 모델의 오분류를 직접 살펴보고, 적대적 공격에도 올바른 판단을 할 수 있는 강인한 모델을 구축하는 것을 목표로 한다.

2. 관련 연구

적대적 공격에 대한 모델의 강인성을 증가시키기 위해 다양한 연구가 진행되었다. Chakraborty 연구팀[5]은 기울기를 이용한 적대적 공격에 대응하기 위해 모델 자체의 기울기를 작게 구성함으로써 적대적 공격

* 교신저자 (corresponding author)



(그림 1) 제안하는 모델 아키텍처

영향을 줄이는 방어적 증류(defensive distillation) 방법을 제안한다. 그리고 이미지에 스무딩 (smoothing) 필터를 적용하거나 픽셀의 색 농도를 줄이는 등의 처리를 통해 적대적 공격에 저항하는 방법인 특징 압축(feature squeezing) 기법을 설명한다.

3. 제안 방법

적대적 공격 대응 방법에는 크게 모델 자체를 강화하는 방법과 이미지 처리를 통한 방어 기법이 있다.

3.1. 모델 강화 기법

적대적 공격에 대해 모델을 강인하게 만드는 대표 기법인 적대적 학습(adversarial training)[3]은 학습 데이터 셋에 적대적 공격이 들어간 이미지를 추가하여 학습을 돌리는 것을 말한다. 신경망은 아래식과 같이 공격받지 않은 일반 이미지와 적대적 예제를 모두 학습에 사용하기 때문에 적대적 공격에 대한 저항성을 갖게 된다. 데이터 로드 후, 학습과정에서 적대적 공격을 시행하여 매 에폭(epoch)마다 새로운 적대적 예제를 생성함과 동시에, 일부 데이터는 공격을 가하지 않도록 설정한다. 지정한 배치(batch) 크기 안에서 일반 이미지와 적대적 예제의 비율을 정해 데이터를 구성하여 학습을 진행한다.

$$\tilde{J}(\theta,x,y) = \alpha J(\theta,x,y) + (1-\alpha)J\left(\theta, x+\epsilon sign\nabla_x J(\theta,x,y)\right)$$

이때, 학습 데이터 셋에서 일반 이미지와 적대적 예제의 구성 비율(α)에 따라 판별 성능의 차이가 나타난다. 두 이미지에 대한 정확도는 상충(trade-off) 관계를 가지는 한계가 있으며, 이를 보완하고자 이미지 처리를 통한 방어 기법도 함께 사용한다.

3.2. 이미지 처리 기반 방어 기법

두 번째 방법으로 이미지 크기 변환(resizing)과 JPEG 압축과 같은 이미지 처리 기반 방어 기법[6]이 있다. 이미지 크기 변환은 지역적 보간(interpolation)을 통해 스무딩 효과를 보이며, 이는 적대적 예제의 섭동을 일부 제거하는 역할을 한다. 크기가 $H \times W$인 적대적 예제를 r배만큼 축소한 뒤, 원래 크기로 다시 확대하는 방식으로 학습 데이터 셋에 적용한다. 적대적 공격에 대한 섭동 제거를 목표로 하는 것이기 때문에 학습과정에서 적대적 예제 생성 후 진행한다.

다음으로 JPEG 압축은 이미지의 고주파 구성요소를 손실시키는 방식으로 작동하는데, 적대적 섭동 또한 고주파에 속하기 때문에 해당 압축을 통해 섭동은 크게 손실된다. JPEG 품질 계수(quality factor)를 조절하며 실험을 진행하였고, 위와 같이 적대적 예제 생성 후에 기법을 적용하였다.

4. 실험 및 결과

4.1. 실험 데이터

오픈소스 데이터 셋인 Celeb-DF[7]와 직접 수집한 이미지로 데이터 셋을 구성한다. Celeb-DF 는 유명인의 실제 영상과 딥페이크가 적용된 영상으로 구성되어 있다. 이때 모든 영상의 길이는 10 초 내외이며, 30 프레임마다 한 장씩 이미지를 추출한다. 이렇게 추출한 이미지에 얼굴 탐지 모델인 MTCNN 을 사용하여 얼굴 부분만 잘라내는 과정을 거친다. 직접 수집한 이미지의 경우에는 별도의 딥페이크 변환기를 적용하여 딥페이크 이미지를 만든 이후에 앞 과정과 동일하게 얼굴 부분을 잘라준다. Celeb-DF 를 통해 8,000 장의 데이터를 수집하고 직접 수집한 이미지로는 1,000 장의 데이터를 모아서 총 9,000 장의 데이터 셋을 구축하였다. 이때, 실제(real) 이미지와 가짜(deepfake) 이미지는 5:5 비율로 구성하였다.

4.2. 실험 구성

실험 구성에는 파이토치(pytorch) 라이브러리를 사용하였고, 한 개의 RTX 3090 그래픽카드를 사용하여 학습과 평가가 이루어졌다. 배치 크기는 128, 최적화 알고리즘으로는 momentum SGD, 손실 함수는 cross entropy loss 를 적용했다. 기반(baseline) 모델은 적대적 학습이나 이미지 처리 기반 방어 기법이 적용되기 전 모델을 말하며, EfficientNet-b0 를 사용했다. 모든 실험에서 30 에폭(epochs) 정도면 충분히 유의미한 결과를 얻을 수 있었다.

4.3. 실험 결과

적대적 학습: 데이터를 로드한 후 적대적 예제를 생성하여 학습을 진행한다. 이미지 전처리에서 정규화를 적용할 경우, 픽셀 값이 [-1,1]로 나오게 된다. 적대적 공격 모듈인 torchattacks 의 입력값은 [0,1]의 픽셀 값으로 맞춰 구현되어 있기 때문에 적대적 공격을 적용한 뒤 정규화를 진행해야 한다. 따라서 모델에 정규화 역할을 수행하는 레이어(layer)를 추가한다. 학습 데이터에서 공격 되지 않은 이미지(original)와 적대적 예제의 구성 비율을 조정하여 학습을 진행하였다. <표 1>을 살펴보면 두 이미지의 비율을 4:6 으로 설정했을 때 비교적 높은 성능이 두 예제에서 고르게 나온다는 것을 알 수 있다.

<표 1> 적대적 예제 비율별 적대적 학습 결과

ori:adv	orignal example		adversarial example	
	loss	acc	loss	acc
(baseline) 10:0	**0**	**100**	4.32	39.1
6:4	0.56	76.7	0.76	54.9
5:5	0.6	75.7	0.68	73.4
4:6	0.63	74.8	0.19	90.4
0:10	11.63	41	**0.07**	**96.1**

<표 2> 이미지 처리 기반 방어 기법 결과

	orignal example		adversarial example	
	loss	acc	loss	acc
baseline (resize 0, jpeg 100)	0	100	4.32	39.1
adversarial training (4:6)	0.63	74.8	0.19	90.4
resize 3	1.05	76.1	0.24	91.3
resize 5	1.58	68	0.38	85.6
jpeg 5	0.68	82.6	0.36	85.2
jpeg 15	0.42	83.4	**0.17**	**93.8**
resize 3, jpeg 15	**0.35**	**85.4**	0.25	91.5

<표 3> 최종 모델 학습 결과

	orignal example		adversarial example	
	loss	acc	loss	acc
Baseline	0	100	4.32	39.1
Adversarial Training (4:6)	0.63	74.8	0.19	90.4
Resize 3, JPEG 15	0.35	85.4	0.25	91.5
FGSM & BIM & PGD	**0.22**	**96.4**	**0.02**	**98.4**

이미지 처리 기반 방어 기법: <표 2>는 이미지 크기 변환과 JPEG 압축을 통한 결과를 나타낸다. 이미지 크기 변환에서는 이미지를 3 배 확대시키는 것이 가장 높은 공격 방어율을 보임을 확인할 수 있다. 또한 Dong 연구팀[6]에서는 JPEG 압축의 경우 압축 계수를 25 미만으로 설정했을 때 높은 방어율을 보인다고 주장하였고, 본 실험을 통해 25 미만의 계수 중 15 일때 가장 좋은 결과를 냈다는 것을 알 수 있다. 또한 선형적 변환인 크기 조정보다 비선형적 변환인 JPEG 압축을 사용했을 때 공격 방어율이 더 높다는 것을 확인해 볼 수 있다.

추가적으로, <표 3>에서 FGSM 뿐만 아니라 BIM, PGD 와 같은 기법을 실험에 모두 적용했을 때의 결과는 단일 방법만 적용했을 때보다 높은 성능을 보임을 알 수 있다.

5. 결론

딥러닝 기반 딥페이크 탐지 모델에 치명적인 오류를 범하게 하는 적대적 공격에 대하여 2 가지 방법을 통해 강인한 모델을 구축한다. 모델을 강하게 만드는 적대적 학습과 이미지 처리 기반 기법을 통해 일반적인 딥러닝 모델과의 확연한 성능 차이를 확인할 수 있다. 이는 딥페이크 뿐만 아니라 AI 부작용 및 위험에 미리 대비하여 범용성을 지닌 알고리즘으로 발전할 수 있으며, 가짜 뉴스 등과 같은 가짜 콘텐츠에 속지 않도록 하여 사회적인 안정성에도 기여할 수 있는 보안 솔루션의 바탕이 된다.

사사

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2022R1A4A1033600). 또한, 본 연구는 2022 년 과학기술정보통신부 및 정보통신기획 평가원의 SW중심대학사업의 연구결과로 수행되었음 (20180002160301001). 추가로, 데이터 셋 구축에 도움을 주신 김지호 님께도 감사의 말씀드립니다.

참고문헌

[1] Kang, Jihyeon, et al. "Detection Enhancement for Various Deepfake Types Based on Residual Noise and Manipulation Traces." IEEE Access 10 (2022): 69031-69040.

[2] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[3] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[4] Lin, Yun, et al. "Threats of adversarial attacks in DNN-based modulation recognition." IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, 2020.

[5] Chakraborty, Anirban, et al. "A survey on adversarial attacks and defences." CAAI Transactions on Intelligence Technology 6.1 (2021): 25-45.

[6] Dong, Xiaoyi, et al. "Robust superpixel-guided attentional adversarial attack." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[7] Li, Yuezun, et al. "Celeb-df: A large-scale challenging dataset for deepfake forensics." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.