[1]적대적 학습을 통한 강건한 딥페이크 탐지 모델 구축

정유나[O1] 배성호[2]
[1]경희대학교 소프트웨어융합학과
[2]경희대학교 컴퓨터공학과
dbsk2480@khu.ac.kr, shbae@khu.ac.kr

Adversarially Robust Deepfake Detection using Adversarial Training

Yuna Jung[O1], Sung-Ho Bae[2]
[1]Department of Software Convergence, Kyung Hee University
[2]Department of Computer Science and Engineering, Kyung Hee University

요 약

딥러닝 기술이 빠르게 발전함에 따라 점점 사람의 눈으로는 구별하기 어려운 딥페이크 영상이 생성되고 있다. 딥페이크가 악용되어 사회적인 피해를 줄 것을 대비하여 딥페이크 영상을 탐지하는 것에 대한 중요성이 커지고 있다. 이에 따라 딥페이크 영상을 탐지하는 다양한 기법들에 대한 연구가 활발히 진행되고 있다. 다양한 탐지 기법 중에서도 Convolutional Neural Network(CNN) 기반의 모델들이 좋은 탐지 성능을 보여주고 있고, 높은 정확도로 진짜 영상과 가짜 영상을 구분하고 있다. 하지만 이 모델들은 적대적 공격(Adversarial Attack)에 취약하다는 단점이 있다. 특히 딥페이크 분야에서 악의적인 목적을 가진 사용자가 충분히 학습된 딥페이크 탐지 모델을 속이는 적대적인 예제들을 생성해낼 가능성이 높기 때문에, 이러한 공격에 강건한 모델을 구축하는 것이 중요하다. 따라서 분류 문제에서 적대적 공격에 대한 가장 강력한 방어 방식으로 알려져 있는 적대적 훈련(Adversarial Training)을 최초로 딥페이크 문제에 적용하여 원본 영상에 대한 정확도도 유지하면서 적대적 공격에도 강건한 모델을 구축하였다.

1. 서 론

인공지능(AI)을 이용해서 타인의 사진을 음란물과 합성하여 음해하거나, 유명인사를 다른 사람과 합성하여 가짜 뉴스를 퍼트려 정치, 사회적으로 피해를 주는 사례가 점점 늘어나고 있다. 이때 인공지능을 기반으로 소스 영상에서 얼굴을 추출하여 타겟 영상에 합성하는 기술을 딥페이크라고 한다. 딥페이크를 생성하는 기술은 전통적인 AE(Auto Encoder) 기반의 생성 방식부터, GAN 이나 그래픽 기반의 생성 방식까지 다양하며, 놀라운 속도로 발전하고 있다. 점점 얼굴 합성 기술이 발전함에 따라서, 사람의 눈으로 진짜 영상과 가짜 영상을 구분하는 것이 어려워지고 있다. 악의적인 목적을 가지고 딥페이크를 이용할 경우, 사회 및 국가적인 혼란을 야기하기 때문에 딥페이크를 탐지하는 연구가 활발히 진행되어 왔다. 그 중에서도 Deep Neural Network 기반의 방식이 다른 탐지 방식들과 비교하여 높은 탐지 정확도를 보인다. 하지만 딥러닝 모델이 적대적 공격에 취약하다고 알려져 있는 만큼, 딥페이크 분야에서 사용하는 DNN 기반의 탐지 모델 또한 적대적 공격에 취약하다는 단점이 있다. 적대적 공격이란 공격자가 사람의 눈으로 확인하기 어려운 작은 섭동(perturbation)을 추가한 입력 데이터로 충분히 학습된 딥러닝 모델을 속이는 방법이다. 이렇듯 충분히 학습되어 높은 정확도로 진짜 영상과 가짜 영상을 구분하는 모델에 작은 perturbation 을 추가한 딥페이크 영상을 넣으면 가짜 영상을 진짜 영상으로 구분하게 된다. 본 논문에서는 딥페이크 분야에서 적대적 공격으로 딥러닝 모델을 속이는 적대적인 가짜 영상을 합성할 가능성이 매우 크다 판단하여, 이를 방어하는 강력한 모델을 만드는 연구를 진행하였다.

단순 분류 문제의 경우, 적대적 훈련(Adversarial Training)이 적대적 공격에 대한 강력한 방어 방식으로 알려져 있다. 하지만 딥페이크 탐지 문제는 단순한 분류 문제보다 훨씬 복잡하기 때문에 아직까지 믿을만한 방어 방식이 존재하지 않는다. 딥페이크 분야에서는 유일하게 [6]에서 방어 방식으로 Deep Image Prior(DIP)와 Lipschitz Regularization 을 제안한다. DIP 는 이미지를 탐지 모델에 넣기 전에 전처리하는 단계에서 적대적 섭동을 제거하고, 원본 이미지로 복원하는 방식이다. 하지만 이 방식의 경우 한 이미지 당 inference 시간이 적어도 30 분 이상이 걸려 실시간으로 활용하기 어렵다는 한계가 있다. Lipschitz Regularization 은 입력에 대해 탐지 모델의 logit 의 gradient 크기에 제약을 주어 작은 perturbation 에 대해 모델이 강건하도록 하는 방식이다. 하지만 이 경우에도 이 방식

[1] "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 사업의 연구결과로 수행되었음"(2017-0-00093)

만을 통한 성능 향상의 폭이 작아 실생활에서 활용하기에는 한계가 있다. 따라서 본 논문에서는 가장 강력한 방어 방식인 적대적 훈련에 Lipschitz Regularization 을 추가하여 높은 정확도로 적대적 공격을 방어하는 강력한 모델을 만드는 방법을 제안한다. 이렇게 학습된 모델은 적대적 샘플뿐만 아니라 원본 영상에 대해서도 어느정도 성능을 유지한다는 장점이 있다. 또한 제안한 방식은 모델이나 데이터에 영향을 받지 않아 다양하게 확장 가능하며, 모델이 추론하는데 추가적인 시간이 들지 않기 때문에 실생활에서도 활용 가능하다는 장점이 있다. 정리하면, 본 논문에서는 적대적 훈련 방식을 통해 적대적 섭동이 추가되지 않은 원본 영상에 대한 탐지 모델의 정확도도 유지하면서, 적대적 섭동이 추가된 샘플에도 강건한 딥페이크 모델을 구축하는 방법을 제안한다. 해당 논문은 강건한 (Robust) 딥페이크 탐지 분야에서는 최초로 적대적 훈련을 사용하였다는 점에서 의의를 갖는다.

2. 제안 방법

2.1. CNN 기반 딥페이크 탐지 모델 학습 방법

딥페이크를 탐지하는 다양한 방법들 중 가장 성능이 좋은 CNN 기반의 모델을 선택하였다. 이 모델은 딥페이크 영상의 각각의 프레임을 독립적으로 진짜 혹은 가짜로 판단하여 최종적으로 해당 영상이 딥페이크 영상인지 아닌지 판별한다. CNN 기반의 딥페이크 탐지 모델을 학습시키는 방법은 다음과 같다:

(1) 모든 딥페이크 영상에 대해서, 영상의 전체 길이를 원하는 프레임 추출 개수로 나누어 각각의 구간별로 프레임을 추출한다.
(2) Face Tracking 모델인 RetinaFace 가 각각의 프레임 내의 인물의 face bounding box 를 찾고, 이를 이용하여 face crop 된 이미지를 생성한다.
(3) crop 된 이미지는 지정한 전처리 방법에 의해 변환되고, 이후 탐지 모델에 입력되어 프레임 별로 독립적으로 진짜인지 가짜인지 판별하도록 학습된다.

2.2. 적대적 샘플 생성 방법

사람의 눈으로는 구분하기 어려울 정도의 적대적 섭동을 추가한 딥페이크 영상을 탐지 모델에 넣었을 때, 대부분의 프레임을 '진짜'로 판단하여 해당 영상을 진짜 영상으로 오분류하도록 하는 것이 적대적 샘플들의 목적이다. 딥페이크 탐지에 자주 사용되는 모델이 정해져 있는 만큼 공격자가 이미지 전처리 방법, 모델 구조, 모델의 파라미터 등 피해 모델에 완벽한 접근이 가능하다고 가정하였다. 적대적 샘플에 대한 딥페이크 탐지 모델의 취약성을 확인하기 위해, 다음과 같은 두가지 공격 방식을 선택하였다: [8] Fast Gradient Sign Method(FGSM) &[9] Projected Gradient Descent(PGD)

2.3. 적대적 샘플에 대한 방어 방법

적대적 샘플들에 대한 방어 방법으로는 분류 문제에서 가장 강력한 방어 방식으로 알려져 있는 적대적 훈련(Adversarial Training)을 최초로 사용하였다. 적대적 훈련이란 모델을 훈련하는 동안 적대적 샘플들을 만들어 모델을 학습시킴으로써, 모델이 적대적 공격에 견고하도록 하는 강력한 방어 방식이다. 본 논문에서는 학습하는 과정에서 적대적 샘플을 만드는 방식으로 PGD 를 사용하였다. 추가적으로 딥페이크 탐지에서 일반화 성능을 높이는데 도움이 된다고 알려져 있는 데이터 증대 방식인 가우시안 노이즈 추가, 가우시안 블러, 이미지 압축 등을 하였고, 기존 손실 함수에 Lipschitz Regularization 을 추가하여 적대적 샘플들에 대한 모델의 견고성을 높였다. Lipschitz Regularization 은 딥페이크 분야에서 소개되었던 적대적 샘플들에 대한 방어 방식으로, 이를 도입한 [6]에 따르면, input 에 대한 탐지 모델 logit 의 gradient 의 크기를 최소화하는 것이 작은 섭동으로부터의 손실을 완화시켜 모델이 적대적 예제에 대해서도 성능을 유지하도록 도움을 준다고 한다. 공식은 아래와 같고, $J_{aug}(x,y,\theta)$는 Lipschitz Regularization 이 추가된 최종적인 손실 함수, $J(x,y,\theta)$는 기존에 학습에 사용되었던 손실 함수, C는 class 개수, N은 입력 벡터의 차원을 의미한다. 또한 $\mathrm{Z(x)}_i$는 i 클래스에 해당하는 logit 값(소프트맥스 계산 전)을, λ는 최종 손실 함수에서의 규제 강도를 조절하는 하이퍼 파라미터를 말한다.

$$J_{aug}(x,y,\theta) = J(x,y,\theta) + \frac{\lambda}{CN}\sum_{i=1}^{C}\|\nabla_x \mathrm{Z(x)}_i\|^2$$

이 방어 방식은 단독으로 사용하기에 성능 향상 폭이 작아 실생활에서 활용하기 어렵다는 단점이 있었는데, 강력한 방어 방식인 적대적 훈련과 결합하여 모델의 강건함을 높일 수 있다.

3. 데이터셋 및 실험 설계

실험에 사용한 데이터셋은 Celeb-DF 이다. Celeb-DF 는 기존 딥페이크 탐지 데이터셋의 품질이 실생활에서 탐지해야하는 딥페이크 영상에 비해 현저히 떨어진다는 문제를 해결하고자 기존보다 향상된 합성 기술을 사용하여 생성된 대규모 데이터셋이다. 이 데이터셋에는 유튜브에 공개되어 있는 성별, 나이, 인종이 다양한 총 59 명의 celeb 들의 590 개의 원본(Real) 영상과 이를 가지고 합성한 5639 개의 딥페이크(Fake) 영상이 존재한다. 또한 유튜브에서 다운받을 수 있는 300 개의 추가 원본(Real) 영상도 제공하고 있다. Celeb-DF 에서 공개한 테스트 셋을 제외한 데이터는 8:2 로 나누어 학습과 검증에 사용하였고, 각각의 영상은 20 개의 프레임을 뽑아서 사용하였다. 또한 딥페이크 영상이 진짜 영상의 개수보다 훨씬 많았기 때문에 진짜 프레임과 가짜 프레임의 개수를 동일하게 맞추어 학습을 진행하였다. 딥페이크 탐지 모델로는 XceptionNet 을 사용하였다. XceptionNet 은 CNN 기반의 모델로, 일반적으로 딥페이크 탐지 분야에서 압축, 또는 압축되지 않은 영상 모두에서 얼굴의 위조된 영역을 식별하는데 가장 성능이 좋은 것으로 알려져 있다.

4. 실험 결과

4.1. 적대적 예제 시각화 및 취약성

<그림 1> 원본 | FGSM | PGD
공격 성공률: | 94.4583 | 99.1733

<그림 1>에서의 공격 성공률은 모델에 적대적 섭동이 추가된 딥페이크 영상을 넣었을 때, 진짜(Real) 영상으로 판별된 비율을 말한다. 원본 영상에 대해서는 97%의 정확도로 가짜 영상을 잘 판별하던 모델에 적대적 샘플을 넣은 결과 FGSM의 경우는 94%, PGD 의 경우는 99%로 가짜 영상을 진짜 영상으로 판별하게 된다.

4.2. 방어 결과

제안했던 적대적 예제에 대한 방어 방식을 크게 세부분—적대적 훈련(AT), 데이터 증강(DA), Lipschitz Regularization(Reg)-으로 나누고, 각 부분들의 여러 조합으로 실험을 진행하며 제안 방식의 효과를 확인하였다. 성능 평가 지표는 정확도(Accuracy)이고, 적대적 샘플(Perturbed)과 원본 샘플(Unperturbed)을 사용하여 해당 모델이 얼마나 적대적 공격에 강건한지와 얼마나 원본 정확도를 유지하는지 모두를 평가하였다. 이때, 적대적 샘플은 가짜 영상에 대해서만 생성하여 모델의 성능을 평가하였다.

Accuracy	Unperturbed		Perturbed	
	Real	Fake	FGSM	PGD
original	95.088	97.0623	5.5417	0.8267
(A)	74.929	88.1016	85.9315	83.8943
(B)	78.3362	92.8993	89.3634	89.6988
(C)	69.0801	94.3313	90.2569	90.7306
(D)	74.2192	96.4275	87.5111	86.4777
(E)	76.9824	**97.6233**	**96.0732**	**95.9061**

<표 1> 적대적 예제에 대한 방어 결과. Original 은 기존 탐지 모델을 학습하는 방식을 의미하고, (A)는 DA+AT (FGSM), (B)는 DA+AT(PGD), (C)는 Random DA(제안했던 세가지 데이터 증강 기법 중 하나만을 랜덤하게 적용하는 방법) + AT(PGD), (D)는 AT(PGD) + Reg, (E) DA + AT(PGD) + Reg 이다.

<표 1>을 보면, (A), (B) 실험을 통해 PGD 를 통한 AT 가 FGSM 을 통한 AT 보다 효과적임을 알 수 있고, (B)와 (C)를 통해 데이터 증대의 효과를 확인할 수 있다. 최종적으로 (E)를 통해서 제안한 방식이 원본 정확도를 유지한 채 적대적 공격에 가장 강건함을 확인하여 제안한 방식이 가장 효과적임을 알 수 있다.

5. 결론 및 한계 분석

CNN 기반의 딥페이크 탐지 모델은 훈련 데이터셋에 대해서 높은 탐지 정확도를 보이지만, 적대적 공격에 취약하다는 단점이 있다. 딥페이크의 경우 충분히 악의적인 사용자가 적대적 샘플을 사용해 잘 학습된 모델을 속일 가능성이 높기 때문에 이에 강건한 모델을 구축하는 것이 중요하다. 따라서 적대적 훈련과 Lipschitz Regularization 을 결합하여 원본 영상에 대한 정확도도 어느정도로 유지하면서 적대적 샘플에 강건한 모델을 구축하였다. 하지만 탐지 모델이 적대적 공격에 강건해질수로 원본 진짜(Real) 영상에 대한 정확도가 떨어졌다. 이는 일반적인 분류 문제에서도 발생하는 적대적 훈련의 근본적인 문제로, 분류보다 훨씬 복잡한 문제인 딥페이크 탐지에서 이 방식을 통해 원본 정확도도 유지하는 동시에 적대적 공격에 강건한 모델을 만드는 것은 어려워 보인다. 따라서 원본 'Real' 영상의 정확도 또한 유지하기 위한 추가적인 연구가 필요하다. 하지만 이 방식은 데이터셋이나 모델에 영향을 받지 않기 때문에, 여러 상황으로 확장이 가능하며, 추론 시간도 늘지 않아 실생활에서 충분히 활용 가능하다는 장점이 있다.

6. 참고문헌

[1] Tolosana, Ruben, Vera-Rodriguez, Ruben, Fierrez, Julian, Morales, Aythami, Ortega-Garcia, Javier. "Deepfakes and beyond: A Survey of face manipulation and fake detection," Information fusion, vol.64, 131-148.

[2] N.Carlini and H. Farid, "Evading deepfake-image detectors with white and black-box attacks," arXiv: 2004.00622, 2020.

[3] P.Neekhara, S. Hussian, M.Jere, F. Koushanfar, and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," arXiv: 2002.12749, 2020.

[4] Zhi Wang, Yiwen Guo, and Wangmeng Zuo, "Deepfake forensics via an adversarial Game," arXiv:2103.13567, 2021.

[5] P. Neekhara, B. Dolhansky, J. Bitton, C.C. Ferrer, "Adversarial threats to deepfake detection: a practical perspective," in *Conference on Computer Vision and Pattern Recognition Workshops* (2021), pp. 923-932.

[6] Apurva Gandhi and Shomik Jain, "Adversarial perturbations fool deepfake detectors," arXiv:2003.10596, 2020.

[7] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Conference on Computer Vision and Pattern Recognition* (2020).

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, arXiv:1412.6572.

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.