적대적 예제를 통한 딥페이크 방지*

정다솔*, 최형기**
성균관대학교

Prevent Deepfake with Adversarial Examples*

Da-sol Jeong*, Hyung-Kee Choi**
Sungkyunkwan Unversity

요 약

딥페이크(Deepfake)란 인공지능(AI)을 사용하여 특정 인물의 얼굴을 다른 사진이나 영상과 합성하여 실제처럼 보이게 하는 기술이다. 이러한 딥페이크 기술이 발전하며 유명인사를 대상으로 했던 딥페이크 음란물 범죄는 일반인들을 대상으로 확대되었다. 그러나 기존의 딥페이크 방어 기술은 딥페이크 탐지에 집중되어 있어 이러한 일반인 대상의 딥페이크 성범죄 피해를 막기 힘들다는 한계가 있다. 이에 본 논문에서는 딥페이크 성범죄를 예방하는 방법으로 심층신경망을 공격하는 공격 기법이었던 적대적 예제를 제시한다. 적대적 예제란 심층신경망의 선형적인 특징을 기반으로 만들어지며, 특정 레이블을 가진 이미지가 해당 레이블로 분류되지 못하도록 만드는 공격이다. 이에 딥페이크와 같은 인공지능이 적대적 예제에 취약한 이유를 설명하고, Cifar-100 데이터셋과 ResNet 모델을 사용하여 사람의 얼굴을 인식하는 심층신경망을 제작한 후, 이를 기반으로 적대적 예제 이미지를 생성한다. 이렇게 만들어진 적대적 예제를 실제 상용화된 어플리케이션에 넣어 적대적 예제가 딥페이크 예방 기술로서 효과가 있는지 확인한다. 연구 결과, 적대적 예제로 만들어진 이미지가 어플리케이션에서 인식되지 않는 것을 확인하였다.

*이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2020R1A2C1012708).

I. 서론

딥페이크란 인공지능(AI) 기술을 사용하여 특정 인물의 얼굴을 다른 사진이나 영상과 합성하여 실제처럼 보이도록 만드는 기술을 의미한다. 이러한 딥페이크 영상은 실제 존재하지 않는 사진 혹은 영상을 만드는 데에 사용되며 일반인의 눈으로는 영상의 진위 여부를 간파하기 쉽지 않다. 기존 딥페이크 성범죄의 피해자는 유명인사가 대부분이었으나 기술의 발전으로 딥페이크 성범죄 피해자는 일반인으로 확장되고 있다. 딥페이크 성범죄에서 영상의 진위여부는 중요하지 않기 때문에, 이러한 딥페이크 성범죄 피해를 최소화하는 방법은 딥페이크 영상 제작을 방지하는 것이다.

이를 방지하기 위해 본 논문에서 제시하는 방법은 '적대적 예제(Adersarial Example)'이다. 원본 이미지에 계산된 작은 왜곡을 추가하면 심층신경망의 이미지 분류 결과를 임의로 바꿀 수 있다. 이러한 왜곡은 아주 작은 값이기 때문에 사람의 눈으로는 원본 이미지와 왜곡이 추가된 이미지를 구분할 수 없지만, 심층신경망은 왜곡이 추가된 이미지를 원본 이미지와 다르게 분류한다. 이렇게 왜곡이 추가된 이미지를 적대적 예제라 하며 심층신경망의 취약점이자 문제점으로 다루어졌다.

그러나 본 논문에서는 적대적 예제를 심층신경망의 공격 기법이 아닌 딥페이크 성범죄의 예방법으로 제시한다. 사람의 얼굴을 인식하는 심층신경망을 제작한 후, 해당 심층신경망을 기반으로 적대적 예제 이미지를 생성한다. 만들어진 적대적 예제를 실제 상용화된 딥페이크 어플리케이션에 넣어 적대적 예제가 딥페이크 예

방 기술로서 효과가 있는지 확인한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개한다. 3장에서는 실험 방법과 실험 결과를 정리하고, 4장에서는 연구 결론 및 의의를 제시한다.

II. 관련 연구

2.1 ResNet

이미지를 분석하기 위해 패턴을 찾는 알고리즘인 Convolutional Neural Network(CNN)의 일종인 ResNet은 이미지 인식과 구분에 사용되는 알고리즘으로 2015년 Kaiming He에 의해 제시되어 ImageNet Large Scale Visual Recognition Challenge(ILSVRC)에서 우승을 차지하며 그 성능을 입증하였다. ResNet은 VGGNet의 구조를 토대로, 기존의 CNN 망과는 달리 입력값을 출력값에 바로 더해줄 수 있도록 지름길(shortcut)을 만들어 학습하는 Residual Block 방식을 사용한다. 일반적인 CNN 심층신경망은 입력이 들어왔을 때 은닉층을 거쳐 출력이 나오는 방식을 사용한다. 이러한 구조에서는 Vanishing Gradient이 발생하여 학습이 제대로 되지 않는 한계점이 있다. ResNet의 Residual Block은 입력이 들어왔을 때 출력으로 연결시키는 방법을 통해 Vanishing Gradient 문제를 해결하여 좋은 성능을 얻을 수 있다.[1]

2.2 적대적 예제

적대적 예제로 인한 심층신경망의 취약성은 2014년 ICLR에서 Szegedy 등에 의해 처음으로 제기되었다. 이들은 원본 이미지에 왜곡을 추가함으로써 적대적 예제를 만들 수 있음을 제시했다.[2] 또한 I. Goodfellow는 적대적 예제의 원인은 심층신경망의 지나친 선형성 때문이라고 주장하며, 이를 기반으로 Fast Gradient Sign Method(FGSM)이라는 적대적 예제 생성법을 만들어냈다.

원본 이미지 x에 대해, 적대적 예제 $x' = x + \eta$에서 η을 만드는 식은 다음과 같다.

$$\eta = \varepsilon \, sign(\nabla_x J(\theta, x, y))$$

적대적 예제는 Loss 함수 그래프의 기울기를 구한 후, x값을 변화시켜 Loss 함수 y값을 크게 만드는 방식이다. 따라서 기울기가 양수면 ε값은 양수가 되고, 기울기가 음수면 ε값은 음수가 되어 η값이 결정된다. 이들은 이러한 방식으로 손실함수의 기울기 부호를 사용하여 왜곡을 추가하는 것만으로 신경망의 예측 오류를 일으킬 수 있음을 증명했다.[3]

2.3 적대적 예제의 전이성

특정 심층신경망을 공격하기 위해 만들어진 적대적 예제는 유사한 구조로 학습된 다른 심층신경망을 공격할 때에도 효과적이다. 이러한 특징을 적대적 예제의 전이성(Transferability)이라고 하며, 전이성의 존재로 인해 적대적 예제에 관한 연구는 심층신경망의 근본적인 문제로 간주되기 시작했다. 적대적 예제가 존재하는 이유와 전이성과 같은 특징을 지닌 이유의 명확한 이론적 규명은 아직 이루어지지 않았다. Szegedy등은 심층신경망의 과적합으로 인한 지나친 비선형성이 적대적 예제의 원인이라고 주장하였으나[2], Goodfellow등은 오히려 심층신경망의 선형적 특징이 적대적 예제의 원인이 된다고 주장하였다.[3] Ilyas등은 적대적 예제는 심층신경망의 오류가 아니라 신경망 분류 성능에 영향을 주는 비강건한(non-robust) 특징이라는 의견을 펼쳤다.[4]

III. 실험 방법 및 실험 결과

3.1 데이터 전처리 및 심층신경망 학습

적대적 예제 이미지를 제작하기 위해, 사람의 얼굴을 인식하는 심층신경망 분류기를 만들었다. 해당 분류기는 많은 수의 클래스 레이블 중 사람의 레이블을 구분하는 업무를 수행해야 하기 때문에, 총 100개의 클래스 레이블에 대해서 예측하며 사람 레이블을 포함하고 있는 Cifar-100 이미지 데이터셋을 사용하여 학습했다. 이미지의 크기는 적대적 예제 생성과 맞춰 224×224로 통일시켰다. 또한 Pytorch 공식 홈페이지에서 이미지를 정규화하기 위해 공식적으로 사용하는 값인 mean=[0.485, 0.456, 0.406] 및 std=[0.229, 0.224, 0.225]를 사용하여 전처리를 진행하였다. 그 후 이미지 인식과 분류에 주로 사용되는 알고리즘인 ResNet-56 알고리즘을 사

용하여 심층신경망을 학습시켰다. 여러 개의 클래스를 구분하는 다중 분류를 위해 CrossEntropyLoss 손실함수를 사용하였으며 최적화로는 SGD를 사용하였다. Batch 사이즈는 32이었으며 심층신경망의 학습 Epoch횟수는 100회 진행하였다.

3.2. 적대적 예제 생성

[그림 1] 적대적 예제 이미지

Epsilon값이 클수록 더 많은 왜곡이 들어갔다.

Epsilon	people image / total iamge	Test Accuracy
0.1	864 / 3295	0.262215478
0.3	698 / 3295	0.211836115
0.7	421 / 3295	0.127769347
1	243 / 3295	0.073748103
1.3	149 / 3295	0.04522003
1.7	78 / 3295	0.023672231
2	47 / 3295	0.014264036
2.3	28 / 3295	0.008497724
2.7	13 / 3295	0.003945372
3	11 / 3295	0.003338392
3.3	10 / 3295	0.003034901
3.7	4 / 3295	0.001213961
4	4 / 3295	0.001213961
4.3	4 / 3295	0.001213961
4.7	2 / 3295	0.00060698
5	2 / 3295	0.00060698

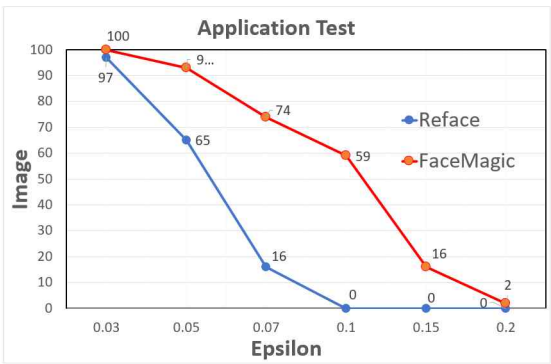
[표 1] 적대적 예제의 심층신경망 테스트 결과

Epsilon 값이 커질수록 Accuracy는 낮아졌다.

심층신경망 학습 결과 테스트와 적대적 예제 생성을 위한 테스트 데이터셋으로는 대표적인 사람 얼굴 데이터셋인 CelebA를 사용하였다. CelebA 데이터셋에서 제공하는 3295개의 사람 얼굴 이미지를 토대로 만들어진 심층신경망을 테스트한 결과, 3295개의 이미지 중 3232개의 이미지가 사람 클래스 레이블에 분류되며 약 98.1%의 정확도를 보였다. 즉, 해당 분류기는 사람의 얼굴 이미지를 제대로 사람 클래스 레이블의 분류하고 있는 것을 확인할 수 있다. 그 후, 만들어진 분류기를 이용하여 FGSM 방법으로 적대적 예제를 생성하였다. 심층신경망 손실함수의 기울기 부호를 구한 후, 그 부호대로 ε 값만큼 왜곡을 발생시켜 이미지에 추가한다. 이러한 방법으로 [그림 1]에서와 같이 적대적 예제를 생성하였다.

ε값은 0.01, 0.03, 0.07, 0.1, 0.13, 0.17, 0.2, 0.23, 0.27, 0.3, 0.33, 0.37, 0.4, 0.43, 0.47, 0.5이었으며, 그 결과는 [표 1]와 같다. ε값이 0.01일 때 3295개의 사람 얼굴 이미지 중 864개만이 사람 클래스 레이블로 분류되며 26%의 정확도를 보였으며, ε값이 점점 커질수록 정확도는 점점 낮아져 ε값이 0.2가 넘은 후에는 정확도가 1% 미만으로 떨어졌다. 즉, ε값이 점점 커질수록 분류기는 사람의 이미지를 사람 클래스 레이블이 아니라 다른 클래스 레이블로 구분하는 것을 확인했다.

3.3 적대적 예제 테스트



[그림 2] 적대적 예제의 어플리케이션 테스트 결과

가로축은 Epsilon값, 세로축은 사람으로 인식된 이미지 개수로 Epsilon값이 커질수록 어플리케이션은 이미지를 인식하지 못했다.

이렇게 만들어진 적대적 예제들을 실제 상용화된 어플리케이션에 넣어 딥페이크 방지 방법으로 효과가 있는지 확인한다. 이러한 방법을 사용할 수 있는 이유는 적대적 예제는 '전이성'이라는 특징을 갖기 때문이다. 특정 심층신경망

을 공격하기 위해 만들어진 적대적 예시는 다른 심층신경망을 공격하는 데에도 효과적이다. 이러한 전이성은 심층신경망이 데이터의 일반적인 특성을 학습했기에 나타난다. 즉, 사람의 얼굴 이미지를 분류하는 심층신경망을 공격하는 적대적 예제는 전이성에 의해 사람 얼굴을 인식하는 딥페이크 어플리케이션에서도 효과를 볼 가능성이 높다.

적대적 예제 테스트를 위해 사용한 어플리케이션은 구글 플레이 스토어에서 높은 다운로드 횟수를 기록한 Reface, FaceMasic이다. Reface는 다운로드 횟수가 1억이 넘은 대표적인 딥페이크 어플리케이션이며, FaceMasic은 다운로드 횟수 100만, 2만개의 리뷰와 4.7의 높은 별점을 받은 딥페이크 어플리케이션이다. 적대적 예제는 CelebA 데이터셋을 사용하여 만들어졌으며, ε값은 0.03, 0.05, 0.07, 0.1, 0.15, 0.2로 각 ε값마다 100장씩의 적대적 예제를 만들어 테스트를 진행하였다. 그 결과는 다음과 같다.

[그림 2]에서와 같이 Reface의 경우 ε값이 0.03일 경우 97장의 얼굴 이미지를 인식했으나, 0.07일 경우 16장의 이미지만을 인식하고 0.1 이상인 경우 단 한 장의 이미지도 인식하지 못하였다. FaceMagic의 경우 ε값이 0.03일 경우 100장의 얼굴 이미지를 모두 인식했으나, 0.1일 경우 59장의 이미지만을 인식하고 0.2일 경우 단 2장만의 얼굴 이미지를 인식하였다. 즉, 어플리케이션마다 차이는 있지만 전체적으로 ε값이 커질수록 낮은 정확도를 보였으며, 적대적 예제 이미지의 얼굴 인식에 실패하여 영상을 제작하지 못하였다.

Ⅳ. 결론

Cifar-100 데이터셋과 ResNet 모델을 기반으로 만들어진 심층신경망은 사람의 얼굴 이미지를 98% 정확하게 라벨링하는 데에 성공하며 높은 정확도를 보였다. 이 심층신경망을 기반으로 FGSM 방법을 사용하여 적대적 예제를 만들었으며, 적대적 예제의 ε값이 점점 커질수록 분류기가 사람의 이미지를 사람 레이블이 아니라 다른 레이블로 구분하는 것을 확인했다. 이렇게 만들어진 적대적 예제를 딥페이크 기술을 사용하여 실제 상용화된 영상 제작 어플리케이션인 Reface와 FaceMagic에 넣어 테스트했다. 그 결과, 사람의 눈으로는 적대적 예제의 이미지를 사람으로 인식할 수 있었음에도 불구하고, 어플리케이션은 얼굴 인식에 실패했다. 즉, 적대적 예제의 원리를 이용하여 이미지에 왜곡을 넣는 방법으로, 사람의 눈으로는 인식 가능하지만 어플리케이션은 인식하지 못하는 딥페이크 방지용 이미지가 만들어진 것으로 볼 수 있다.

[참고문헌]

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition” 2015

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, “Intriguing properties of neural networks.”, the International Conference on Learning Representations (ICLR), 2014

[3] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and harnessing adversarial examples.”, the International Conference on Learning Representation (ICLR), 2015

[4] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” Neural Information Processing Systems (NeurIPS), 2019