캡스톤 디자인
‘딥페이크 탐지’

#3. 딥페이크 탐지 우회 기술 관련 논문 조사

김지수, 김민지, 민지민

Anti-Forensics (안티 포렌식)

배경

: **디지털포렌식** 기술의 발전

→ **디지털 증거물의 법적 효력** 중요 → 이를 **회피하려는 움직임**

안티포렌식

: 디지털포렌식 기술에 대응해 **데이터를 조작, 삭제, 또는 난독화**하여 자신에게 불리하게 작용할 수 있는 **증거의 존재, 또는 그 양과 질에 부정적인 영향**을 미치거나 **증거분석을 곤란**하게 하는 행위

Deepfake-Image Anti-Forensics with Adversarial Examples Attacks

게재된 저널 : Future Internet

발행일 : 2021. 11. 17

링크: https://www.mdpi.com/1999-5903/13/11/288

배경: 딥페이크 탐지기의 대부분은 **adversarial 사례에 대해서 취약**하다. 따라서 딥페이크 탐지기를 개선하기 위해 **adversarial 사례의 영향**을 알아야 한다. → 딥페이크 탐지기에 대한 **안티포렌식 사례 연구** 개발

방법: Poisson noise DeepFool(PNDF)라는 반복적인 adversarial 사례 생성 방법을 사용해 다른 direction의 **이미지에 perturbation을 추가**해서 딥페이크 탐지기의 정확도를 줄인다.

Deepfake-Image Anti-Forensics with Adversarial Examples Attacks

- **Data Perturbation**: 데이터의 작은 변화

  → 데이터에 **기밀성을 부여**하기 위해 **'노이즈'를 추가**하는 보안 기법

- **Poisson Noise DeepFool(PNDF)**

  → 예측을 변화시킬 수 있는 **최소한의 perturbation 진행**

  → poisson noise는 매우 작고, 정규 poisson distribution을 따름

  → 신호 의존성과 신호대잡음비를 개선하는 우수한 성능 때문에 directional perturbation을 진행하면서 노이즈를 증가시킬 때 이미지의 품질 보장 가능

Deepfake-Image Anti-Forensics with Adversarial Examples Attacks

실험 데이터셋 생성

→ 다양한 GAN모델 사용

Table 1. Generation Models Datasets.

Category	Models	Fake Images	Real Images	Image Source
Conditional GANs	CycleGAN	1321	1321	Style/object transfer
	StarGAN	1999	1999	Celebes
	GauGAN	5000	5000	COCO
Unconditional GANs	ProGAN	4000	4000	LSUN
	StyleGAN	5991	5991	LSUN
	Bagan	2000	2000	ImageNet
Low-level vision	SITD	180	180	Raw camera
	SAN	219	219	Standard SR benchmark
Perceptual loss	CRN	6382	6382	GTA
	IMLE	6382	6382	GTA
DeepFakes	FaceForensics++	2700	2700	Videos of faces

perturbation 전후 실험 결과

→ **perturbation 적용 후**

탐지 성능 대폭 하락

Table 3. Generalization comparison of the detector before and after adding perturbation by PNDF.

Models	f-ac		AP		AUC	
	Unperturbed	Perturbed	Unperturbed	Perturbed	Unperturbed	Perturbed
CRN	0.9987	0.0007	0.9823	0.3135	0.9877	0.0394
IMLE	0.9976	0.0393	0.9840	0.3069	0.9884	0.0024
SITD	0.8666	0.0611	0.9723	0.3068	0.9756	0.0156
StarGAN	0.8129	0.0065	0.9400	0.3126	0.9780	0.0417
CycleGAN	0.7887	0.0832	0.9246	0.3346	0.9384	0.0799
StyleGAN	0.7426	0.1330	0.9959	0.3100	0.9958	0.0354
GauGAN	0.6480	0.0180	0.8948	0.3077	0.9206	0.0130
BigGAN	0.4690	0.0465	0.8450	0.3087	0.8819	0.0160
DeepFake	0.0685	0.0111	0.8902	0.3063	0.8844	0.0000
SAN	0.0182	0.5844	0.7046	0.3400	0.7185	0.1049

Evading Deepfake-Image Detectors With White- and Black-Box Attacks

게재된 저널 : IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020

발행일 : 2020

링크:
https://openaccess.thecvf.com/content_CVPRW_2020/papers/w39/Carlini_Evading_Deepfake-Image_Detectors_With_White-_and_Black-Box_Attacks_CVPRW_2020_paper.pdf

배경: 딥페이크 탐지기가 다양한 adversarial 공격에 취약하다. 단일 생성기에 한해 AUC 0.95에 달하는 탐지기가 취약해지는 5개의 공격 사례 연구를 개발한다.

방법

- (**화이트박스 공격**) Distortion-minimizing Attack
- (화이트박스 공격) Loss-Maximizing Attack
- (화이트박스 공격) Universal Adversarial-Patch Attack
- (화이트박스 공격) Universal Latent-Space Attack
- **블랙박스 공격**

Evading Deepfake-Image Detectors With White- and Black-Box Attacks

화이트 박스 공격

- **Distortion-minimizing Attack**: 가장 낮은 차수의 비트를 플립하고, 플립하는 픽셀의 비율 최소화

- **Loss-Maximizing Attack**: 합성 이미지가 탐지되지 않을 가능성을 최대화하는 목적함수 정의

→ 이미지 영역의 1% perturbation

- **Universal Adversarial-Patch Attack**:잘 탐지하지 못하도록 fake 이미지에 중첩될 patch생성

-**Universal Latent-Space Attack**: GAN모델을 수정해서 adversarial이미지를 생성하는 잠재 공간 공격

→ synthesizer의 잠재 공간에 노이즈 패턴 추가

Evading Deepfake-Image Detectors With White- and Black-Box Attacks

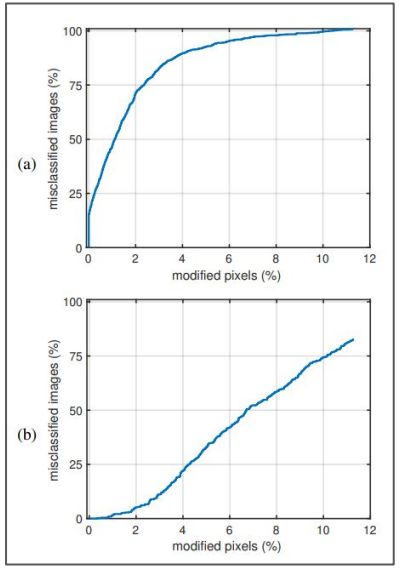
블랙 박스 공격

- 상대가 탐지기의 세부사항에 접근할수 없지만 어떤 유형의 탐지기가 있는지 알고 있는 상황에 사용

<화이트박스 공격에 대한 결과>

(a) → fake image as real

(b) → real image as fake



Dodging DeepFake Detection via Implicit Spatial-Domain Notch Filtering

게재된 저널 : arXiv Preprint, 2021

발행일 : 2021. 11. 25

링크: https://arxiv.org/pdf/2009.09213.pdf

배경

- 딥페이크 기술이 매우 발전했다. 이에 따라 딥페이크 탐지를 회피하는 기술 발전이 궁극적으로 딥페이크 탐지 발전에 도움이 될 것이다.

방법

- **Implicit Spatial-Domain Notch Filtering**을 통해 이미지 품질을 손상시키지 않고 **fake이미지의 artifact 패턴을 줄임**

Dodging DeepFake Detection via Implicit Spatial-Domain Notch Filtering

Notch Filter: 특정 주파수 대역의 성분만을 제거하는 필터

→ 공간영역의 주기적인 노이즈를 제거하는 데 효과적

→ 수동 설계가 필요해서 실행할 수 없음

→ 논문에서는 학습 기반 접근법에 의존하지만, 공간 영역에만 의존하는 방식으로 진행

DeepNotch

- 주기적인 노이즈 패턴을 깨기 위해 **방대한 공간 노이즈 추가**
- 노이즈 없는 fake 이미지 재구성하기 위해 deep image filtering 사용

Deep Image Filtering

- 노이즈가 많은 이미지 픽셀마다 특수 필터 제공 → 딥페이크보다 잘 필터링된 이미지 생성
- 이미지 의미 정보를 사용해서 지능적 노이즈 추가

—> 기존 탐지기의 정확도를 평균 36%, 최대 97% 감소 시킴

FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction

게재된 저널 : arXiv Preprint, 2020

발행일 : 2020년 8월 17일

링크 : https://arxiv.org/abs/2006.07533

-배경 :

현재 GAN 기반 이미지 생성 방법은 아직 불완전하며, 업샘플링 설계는 합성 이미지에 특정 아티팩트 패턴을 남기는 데 한계

-방법 :

이미지 합성 중에 유입되는 **아티팩트를 효과적이고 효율적으로 줄이기 위해** 학습된 선형 사전을 통해 가짜 이미지의 얕은 재구성을 수행하는 **FakePolisher라는 단순하면서도 강력한 접근 방식을 고안**

세 가지 최첨단 가짜 이미지 탐지 방법의 정확성을 크게 감소시킴. 평균 47%, 최악의 경우 최대 93%까지 감소

FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction

<3 단계로 구성 >

1. 먼저 실제 이미지의 패턴을 포착하기 위해 사전 모델을 훈련. → 본질적으로 저차원적인 하위 공간을 형성하며, 실제 이미지의 필수 구조와 표현을 압축적으로 캡처함.
2. 이 사전을 기반으로 사전의 과잉 완성도에 따라 선형 투영 또는 희소 코딩을 통해 앞서 언급한 하위 공간을 사용하여 DeepFake 이미지의 표현을 모색.
3. 이러한 표현을 얻으면 해당 사전을 사용하여 딥페이크 이미지의 '페이크 프리' 버전을 얕게 재구성하여 딥페이크가 도입하는 아티팩트 패턴을 크게 줄일 수 있음.

Adversarial Threats to DeepFake Detection: A Practical Perspective (DeepFake 탐지에 대한 적대적 위협: 실용적인 관점)

게재된 저널 : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops

발행일 : 2021년

내용 : DeepFake 탐지기에 대한 적대적 공격

적대적 변화(perturbation)가 다른 모델에 걸쳐 전달되는 정도와 적대적 사례의 전달성을 개선하기 위한 기술을 제안한다. 또한 실현 가능한 공격 시나리오를 제시하는 **범용 적대적 변화(universal adversarial perturbation)을 사용하여 더 접근하기 쉬운 공격을 만든다.**

https://openaccess.thecvf.com/content/CVPR2021W/WMF/html/Neekhara_Adversarial_Threats_to_DeepFake_Detection_A_Practical_Perspective_CVPRW_2021_paper.html

Adversarial Threats to DeepFake Detection: A Practical Perspective (DeepFake 탐지에 대한 적대적 위협: 실용적인 관점)

방법 :

1. **Threat model** : 시스템 위협을 올바르게 식별하고 해결하면서 체계적인 접근 방식을 허용하는 컴퓨터 보안 최적화 프로세스이다. 보안 사양을 개발하고 해당 사양의 무결성을 테스트 하는 것이다.**시스템 위협 및 취약성을 식별하기 위해 공격자가 사용하는 동기 및 방법을 정확히 파악하는 데 필요**하다.

2. **Victim model** : **프레임 분류 문제를 추가로 분해**한다. 얼굴 추적 모델은 바운딩 박스 감지한다. 자른 면은 일부 입력 변환 및 CNN 분류의 입력으로 제공된다.

3. **White box model** : **공격자가 모든 정보를 알고 공격하는 것**을 의미한다. 적대적 공격을 수행할 수 있는 공격 표면이 노출되게된다.

4, **Black box model** : **공격자가 모델에 대한 정보를 알지 못하고 공격하는 것을 의미**한다. 사전에 제작된 값을 입력하여 출력되는 결과를 관찰하여 모델의 취약성 분석

5. **Universal attack** : **Deepfake 탐지기를 쉽게 속일 수 있도록 더 쉽게 접근할 수 있는 적대적 공격을 설계하는 것을 목표**로 한다.

A CycleGAN Adversarial Attack Method Based on Output Diversification Initialization (출력 다양화 초기화에 기초한 CycleGAN 적대적 공격방법)

게시된 저널 : Journal of Physics

발행일 : 2021

배경 : **생성적 적대 네트워크(GAN)의 강력한 이미지 생성 기능은 이미지와 관련된 애플리케이션에 큰 위협을 가한다.** 위조 이미지를 탐지하는 기술적 방법이 등장했는데, 이는 이후에 교정 조치를 촉발할 수는 있지만 악의적으로 변조된 콘텐츠가 네트워크 미디어를 통해 퍼지는 것을 막을 수는 없다.

내용 : **출력 다양화 초기화에 기초한 적대적 공격방법**

적대적 공격의 시각적 효과를 개선하기 위해, **CycleGAN의 출력 다양화 초기화(ODI)에 기반한 적대적 공격 알고리즘을 제안**. 먼저 출력 다양화 초기화를 활용하여 **적대적 공격에 대한 효과적인 시작점을 찾은 다음, PGD(Project Gradient Descent)를 사용하여 적대적 손실 함수를 수정하여 스타일 전송 네트워크를 반복적으로 공격**한다. 실험 결과는 ODI의 도입이 적대적 출력과 원래 출력 사이의 거리를 효과적으로 확대할 수 있다는 것을 보여준다.

https://iopscience.iop.org/article/10.1088/1742-6596/1948/1/012041/meta

A CycleGAN Adversarial Attack Method Based on Output Diversification Initialization (출력 다양화 초기화에 기초한 CycleGAN 적대적 공격방법)

Cycle GAN??

- 사람이 눈으로 보고 이해할 수 있는 어떠한 이미지 내의 도메인을 스스로 학습하여, 서로 페어하지 않은 이미지끼리라도 학습하여 변형하고 적용이 가능하도록 만든 모델입니다.

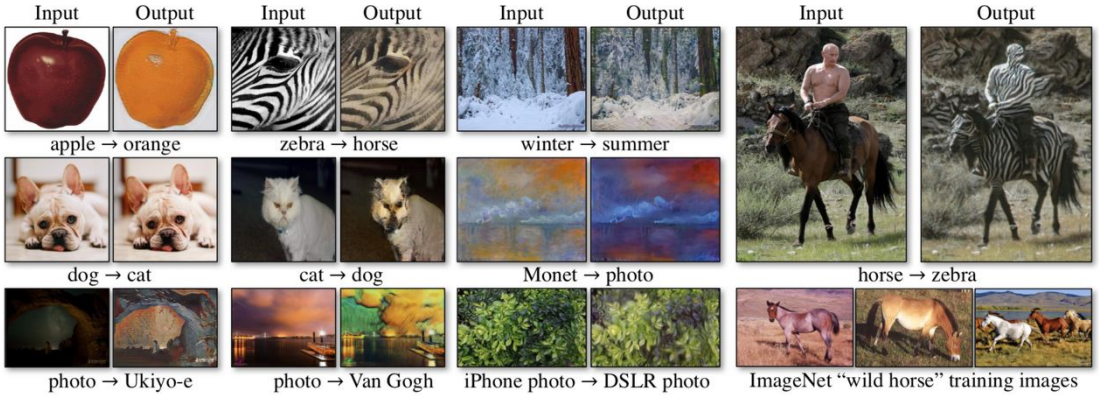


Figure 17: Typical failure cases of our method. Left: in the task of dog→cat transfiguration, CycleGAN can only make minimal changes to the input. Right: CycleGAN also fails in this horse → zebra example as our model has not seen images of horseback riding during training. Please see our website for more comprehensive results.

A CycleGAN Adversarial Attack Method Based on Output Diversification Initialization (출력 다양화 초기화에 기초한 CycleGAN 적대적 공격방법)

방법 :

분류 모델에 대한 적대적 공격의 경우, ODI는 원래 입력에 근접하게 할 수 있는 무작위 초기화 방법을 대체하기 위해 제시된다. 따라서, 적대적 사례를 구성하기 위한 **초기화 방법으로 ODI를 도입**한다. **출력 공간의 거리를 기준으로 원래 입력에서 최대한 멀리 떨어진 출발점을 직접 생성하고, 반복에 대한 그라데이션 기반 방식을 적용해 출력 이미지가 원본 이미지와 확실히 다른지 확인한다**.

A CycleGAN Adversarial Attack Method Based on Output Diversification Initialization (출력 다양화 초기화에 기초한 CycleGAN 적대적 공격방법)

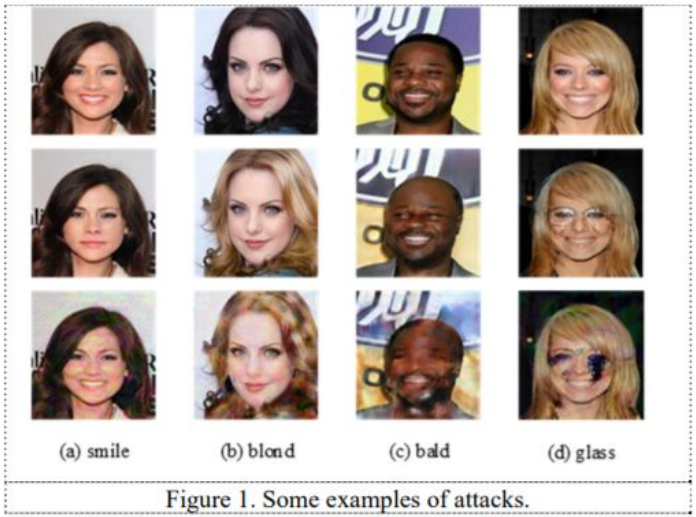


Figure 1. Some examples of attacks.

CycleGAN의 역할에 따라 웃음을 짓거나 금발이거나 대머리거나 안경의 유무에 대한 공격을 할 수 있다.
위 사진을 보면 이미지에 대한 적대적인 공격을 성공적이었다고 볼 수 있다

Anti-Forensics for Face Swapping Videos via Adversarial Training (적대적 훈련을 통한 비디오 얼굴 스와핑 방지)

게재된 저널 : IEEE Transactions on Multimedia

발행일 : 2021년 7월 26일

내용 : **안티포렌식 도구로 동작하는 GAN 모델을 제안, 손실 함수는 제안된 모델의 효율성을 향상시키도록 설계**

실험에 의해 평가된 결과, 우리는 DeepFake 포렌식 탐지기가 제안된 방법에 의해 시작된 공격에 취약하다는 것을 보여준다. 다른 반포렌식 접근법과 비교했을 때, 이것은 딥페이크 반포렌식스가 이룬 엄청난 발전이다. 우리가 제안한 방법으로 **시작된 공격은 탐지 알고리즘과 사람의 눈을 동시에 속일 수 있기 때문에 진정한 딥페이크 안티 포렌식으로 볼 수 있다.**

링크:https://ieeexplore.ieee.org/abstract/document/9495178?casa_token=iBhVgQgmM6cAAAAA:sosdkjxCjddpJKHZeqvkw24nQ9z6w9GqHjfDof27hv4hrg_92cefcAAnyygfbxMPhq08So-y

MagDR: Mask-Guided Detection and Reconstruction for Defending Deepfakes (MagDR: 딥페이크 방어를 위한 마스크 유도 탐지 및 재구성)

게재된 저널 : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

발행일 : 2021년

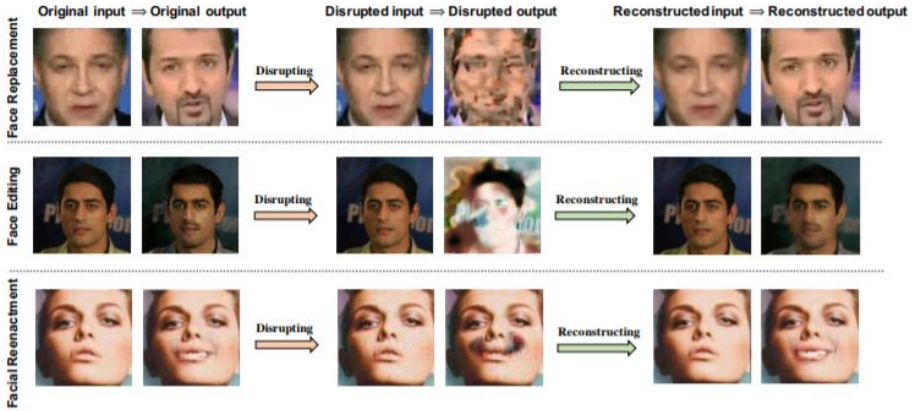
배경 : 딥페이크는 비주얼 콘텐츠의 진위성에 심각한 우려를 제기했다. 소스 데이터에 적대적 변화를 추가하여 딥페이크를 교란할 가능성을 밝혔지만, 아직 위협이 제거되지는 않음

내용 : **적대적 공격으로부터 딥페이크를 방어하기 위한 마스크 유도 탐지 및 재구성 파이프라인인 MagDR을** 제시한다. MagDR은 딥페이크 출력의 이상을 판단하기 위한 몇 가지 기준을 정의하는 검출 모듈로 시작한 다음 학습 가능한 재구성 절차를 안내하는 데 사용한다. 적응형 마스크는 국소 안면 영역의 변화를 포착한다.

실험에서 MagDR은 **딥페이크의 (얼굴 교체 (상단), 얼굴 편집 (중간) 및 얼굴 재현 (하단)) 세 가지 주요 작업을 방어**하고 학습된 재구성 파이프라인이 입력 데이터에 걸쳐 전송되어 **블랙박스 공격과 화이트박스 공격**을 모두 방어하는 데 유망한 성능을 보여준다.

링크:https://openaccess.thecvf.com/content/CVPR2021/html/Chen_MagDR_Mask-Guided_Detection_and_Reconstruction_for_Defending_Deepfakes_CVPR_2021_paper.html

MagDR: Mask-Guided Detection and Reconstruction for Defending Deepfakes (MagDR: 딥페이크 방어를 위한 마스크 유도 탐지 및 재구성)



MagDR은 마스크 유도 감지 및 재구성을 의미

검출기와 재구성기의 두 가지 주요 구성요소가 포함되며,

출력 이미지에서 적대적 공격의 존재를 감지하고 모든 기준을 허용 가능한 값으로 억제하도록 조정 가능한 알고리즘을 수행하는 것

정확한 기준과 더 강력한 검출기 및 재구성기를 사용해서 MagDR의 성능을 향상시킬 수 있다.

Defending against GAN-based DeepFake Attacks via Transformation-aware Adversarial Faces(변환 인식 적대적 얼굴을 통한 GAN 기반 DeepFake 공격 방어)

게재된 저널 : IEEE 2021 International Joint Conference on Neural Networks (IJCNN)

발행일 : 2021년 9월 20일

배경 : 페이스 스왑으로 인해 가짜 콘텐츠(예: 이미지, 비디오)는 인간이 감지할 수 없게 되었다. 딥페이크 공격에 대한 다양한 탐지 기술이 탐구되었다. 그러나 이러한 방법은 고품질 가짜 콘텐츠가 생성된 후 완화 전략이기 때문에 DeepFake에 대한 소극적인 조치이다.

내용 : 우리는 생성 중 차별화된 무작위 이미지 변환을 활용하는 **GAN 기반 DeepFake 공격에 대한 방어책으로 새로운 변환 인식 적대적 교란 얼굴을 사용할 것을 제안한다.**

또한 블랙박스 설정에서 GAN 기반 DeepFake 변형에 대한 **방어 견고성을 강화하기 위한 앙상블 기반 접근 방식을 제안**한다. 적대적 얼굴을 가진 DeepFake 모델을 훈련하면 합성된 얼굴의 품질이 크게 저하될 수 있음을 보여준다

링크:https://ieeexplore.ieee.org/abstract/document/9533868?casa_token=q1V_jifDYR0AAAAA:_bkrq5ZkN-m2QNoBMKPNqT-WUOZEFX0ddM4pzFT5mfu_zFL-NK2qT1-FeZyV3QCFxs78g_cl

Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples

게재된 저널 : IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

발행일 : 2021년

링크:
https://openaccess.thecvf.com/content/WACV2021/html/Hussain_Adversarial_Deepfakes_Evaluating_Vulnerability_of_Deepfake_Detectors_to_Adversarial_Examples_WACV_2021_paper.html

– 배경 :

가짜 비디오 생성은 잠재적으로 악의적인 목적으로 사용될 수 있으므로 Deepfake 탐지기의 적대적 입력에 대한 취약성을 해결하는 것이 중요

–방법 :

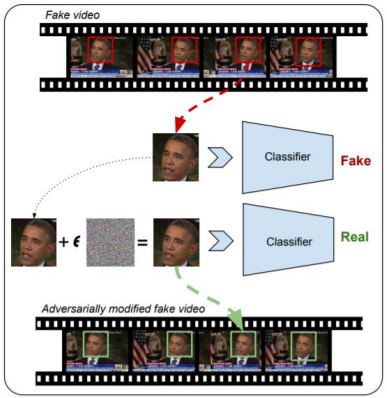
본 연구에서는 기존의 딥페이크 생성 방법을 사용하여 합성된 **가짜 비디오를 적대적으로 수정하여 이러한 탐지기를 무시**할 수 있음을 보여줌.

이 방법은 가짜 비디오를 생성하기 위한 기존 기술을 강화하여 주어진 가짜 비디오 탐지기를 우회할 수 있음.

Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples

- ➔ 즉, 기능적으로 조작된(가짜) 비디오 입력과 Deepfake 검출기가 주어지면, 우리의 작업은 **가짜 비디오를 적대적으로 수정하여 대부분의 프레임이 Deepfake 검출기에 의해 Real로 분류되도록** 하는 동시에 **적대적 수정을 거의 감지할 수 없게 하는 것** .
- ➔ 적대적 예로는 분류 모델을 속일 수 있는 의도적으로 교란된 입력이 있고, 본 연구에서는 FaceSwap(FS), Face2Face(F2F), DeepFake(DF) 및 NeuralTextures(NT)를 사용하여 합성된 가짜 비디오에 대한 적대적 예를 생성.
- ➔ 또한 본 연구에서는 적대적 동요를 제한하기 위해 L1 왜곡 메트릭을 사용하고 경사 부호 기반 방법을 사용하여 이를 최적화

Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples



< 적대적 가짜 비디오를 생성하는 과정>

- 먼저, 주어진 프레임에 대해서, 자른 면에 대한 적대적인 예를 만들어 일부 이미지 변환(정규화 및 크기 조정)을 거친 후 분류기에 의해 Real로 분류됨.
- 그런 다음 적대적인 얼굴은 원래 프레임에서 페이스 크롭의 경계 상자에 배치되고, 이 프로세스는 비디오의 모든 프레임에 대해 반복되어 적대적으로 수정된 가짜 비디오를 만들게 됨.

Exploring Adversarial Fake Images on Face Manifold

게재된 저널 : IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

발행일 : 2021년

링크:
https://openaccess.thecvf.com/content/CVPR2021/html/Li_Exploring_Adversarial_Fake_Images_on_Face_Manifold_CVPR_2021_paper.html

- 배경 :

딥 페이크의 생성과 확산을 방지하기 위해 기존 포렌식 모델의 견고성과 일반화 능력을 향상시키기 위한 안티 포렌식 작업을 개발하고 연구할 필요가 있음

-방법 :

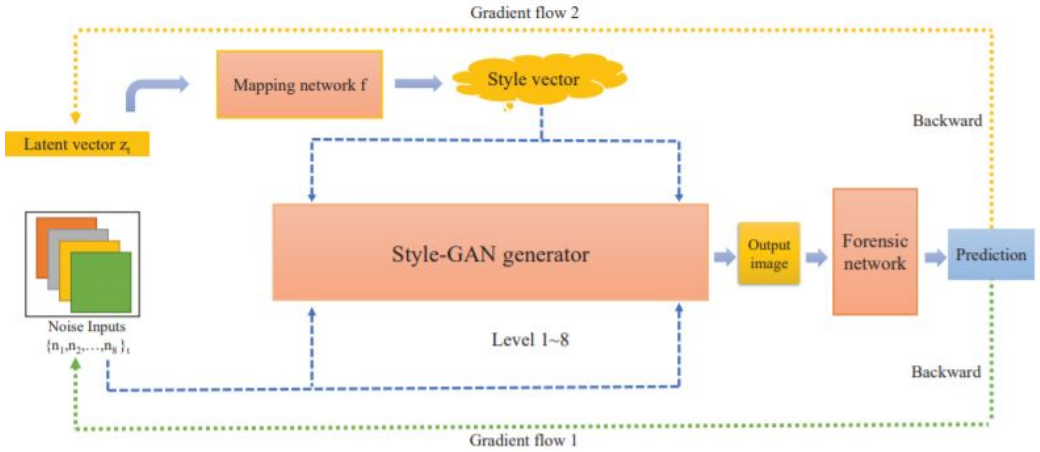
적대적 노이즈를 추가하는 대신 **얼굴 매니폴드에서 적대적 포인트를 최적으로 검색하여 안티포렌식 가짜 얼굴 이미지를 생성**

이 방법으로 생성된 이미지는 Xception 및 EfficientNet의 두 가지 이미지 포렌식 모델을 성공적으로 우회할 수 있음.

Exploring Adversarial Fake Images on Face Manifold

- ➔ Style-GAN의 원래 아키텍처를 변경하지 않고 얼굴 매니폴드에서 적대적 이미지를 찾기 위해 주어진 벡터 주변에서 **Style-GAN의 입력 잠재 벡터 z 또는 노이즈 벡터 n을 조작**합니다.
- ➔ **Style-GAN의 노이즈와 잠재 벡터에 경사 하강법을 적용**하여 포렌식 모델의 예측 손실을 최대화하는 방향으로 벡터를 업데이트하는 것

Exploring Adversarial Fake Images on Face Manifold



< 전체 파이프라인 >

매핑 네트워크 f와 발생기의 업샘플링 레이어의 각 레벨에 삽입된 노이즈 입력 n1, n2...n8 이전의 잠재 벡터 z에 대한 경사 강하를 수행합니다. 그리고 매번, 손실 함수의 한 단계 기울기 기호를 사용하여 고정된 단계 크기 하이퍼 파라미터로 잠재 벡터를 업데이트합니다.