캡스톤 디자인
'딥페이크 탐지'

#1. 데이터셋 정보, 개발물, 논문 정보 조사

김지수, 김민지, 민지민

데이터셋 정보

1. 딥페이크 변조 영상 데이터

- 특징
  - 데이콘 **‘딥페이크 변조 영상 탐지 AI 경진대회’** 데이터셋
    https://dacon.io/competitions/official/235655/data (링크)
  - AIHUB ‘비전’ 카테고리에도 존재
  - 원본 영상에 대해 **GAN기반 변형 알고리즘**을 통해 생성
  - 640828장의 **real 이미지**, 787748장의 변조된 **fake 이미지**로 구성
  - 1920*1080 크기의 이미지

데이터셋 정보

1. 딥페이크 변조 영상 데이터

- 예시

<fake>

<real>

데이터셋 정보

2. 안면 인식 영상

- 특징
  - **AIHUB '안전'** 카테고리 데이터셋
  - https://aihub.or.kr/aidata/34143
  - 얼굴 인식 분야의 위변조 방지를 위한 인공지능 모델 학습용 데이터
  - **프린트, 영상 리플레이, 3D 마스크**를 사용한 얼굴 위변조 공격 데이터를 일반 RGB카메라를 통해 확보
  - 54,870개의 **영상(원본+변조),** 1,646,100개의 **이미지 (동영상 당 30개의 이미지 추출)**

데이터셋 정보

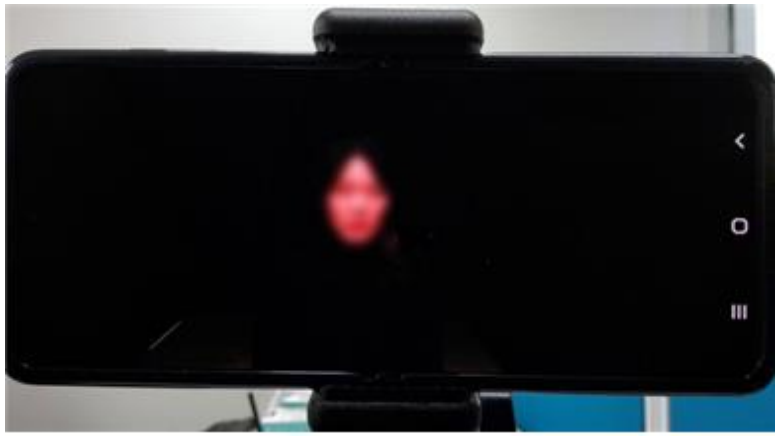
2. 안면 인식 영상

- 예시

<Live 원본>

<Print Attack>

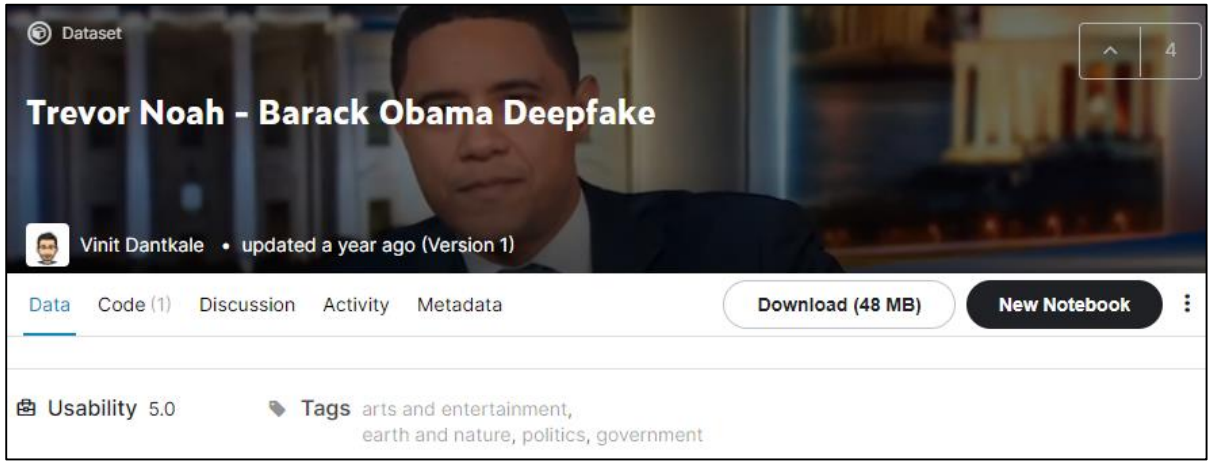
<Replay Attack>



데이터셋 정보

3. Trevor Noah - Barack Obama Deepfake

- 특징
  - **Kaggle** https://www.kaggle.com/vinitdantkale/deepfake?select=fakevideo.mp4
  - 4개의 영상으로 구성
  - real영상, 합성된 fake영상, real영상과 fake영상을 가로로 붙인 영상, 세로로 붙인 영상
  - License 불명확



데이터셋 정보

3. Trevor Noah - Barack Obama Deepfake

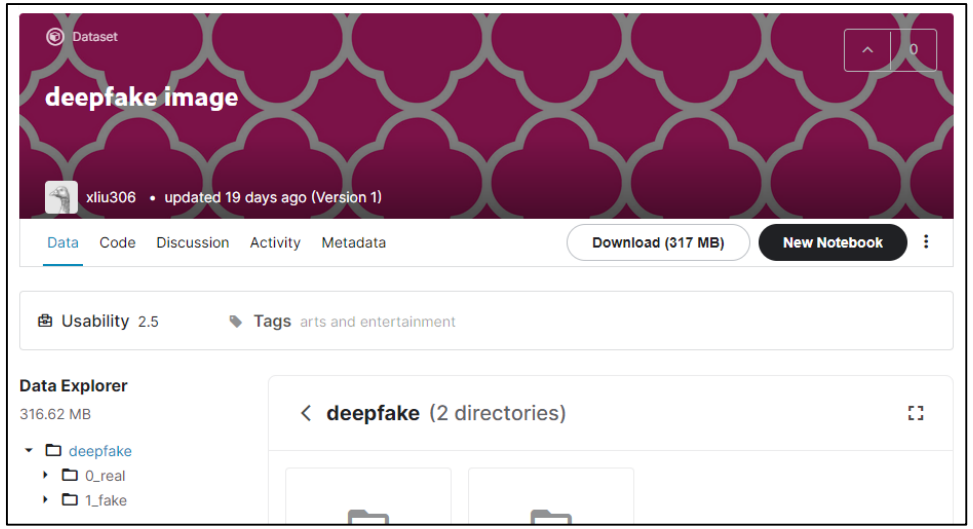
- 예시

<real>

<fake>

데이터셋 정보

4. Deepfake image



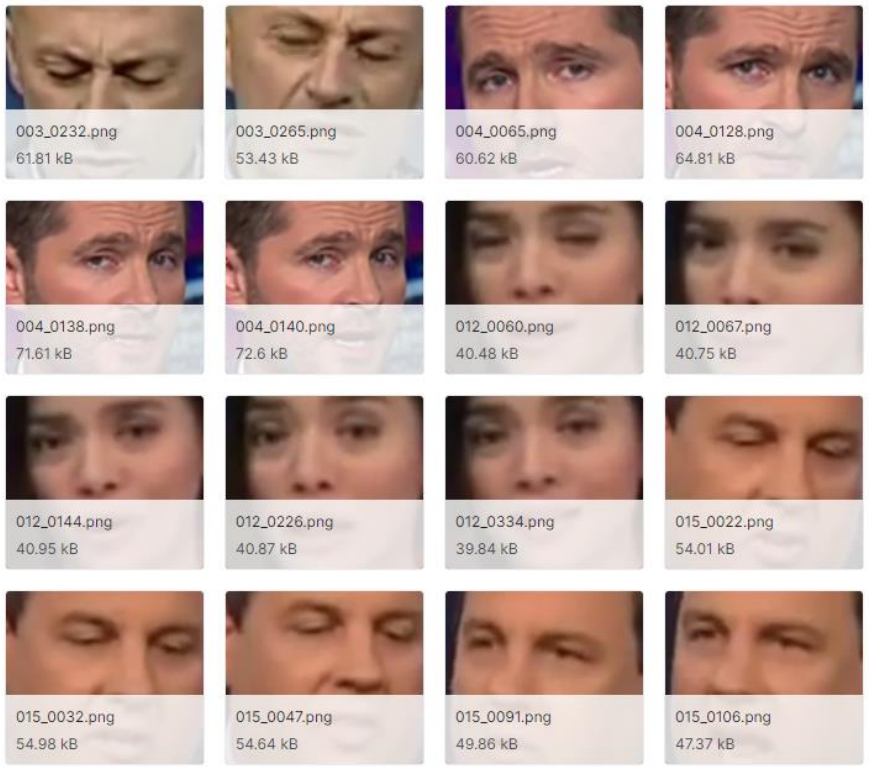
- 특징
  - **Kaggle** https://www.kaggle.com/xinboliu404/deepfake-image
  - **모두 이미지로 구성**되었고, 이미지 크기가 일정하지 않음
  - 2707장의 real이미지와 2698장의 fake이미지로 구성
  - License 불명확

데이터셋 정보

4. Deepfake image

- 예시

<real>

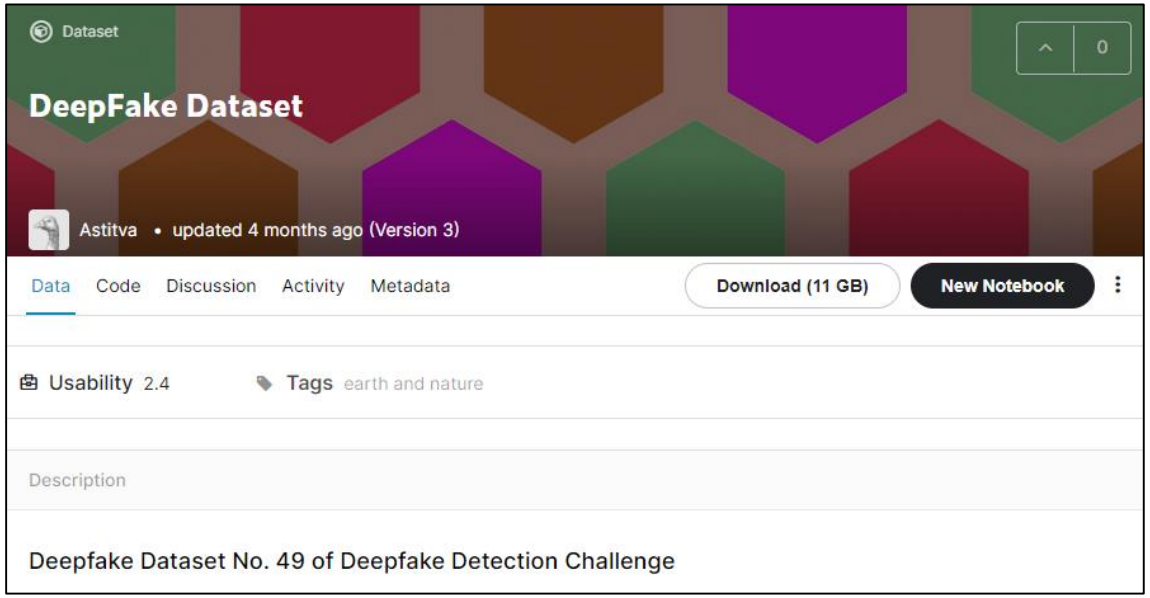


<fake>

데이터셋 정보

5. Deepfake Dataset

- 특징
  - **Kaggle** https://www.kaggle.com/astitvas/deepfake-dataset
  - 3135개의 영상으로 구성
  - 모두 10초의 영상
  - real/fake 구분되어있지 않음
    - License 불명확



데이터셋 정보

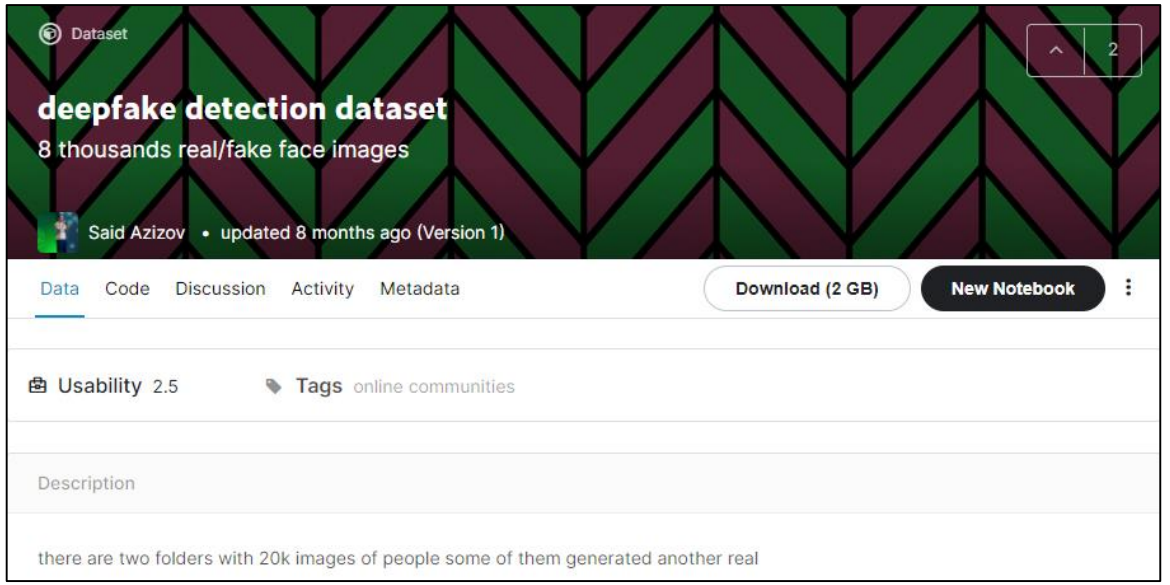
5. Deepfake Dataset

- 예시

데이터셋 정보

6. Deepfake detection datset

- 특징
  - **Kaggle** https://www.kaggle.com/michaelcripman/deepfake-detection-dataset
  - 8000개의 train이미지, 12,000개의 test이미지로 구성
  - real/fake 구분되어있지 않음
    - License 불명확



데이터셋 정보

6. Deepfake detection datset

- 예시

데이터셋 정보

7. ffhq-dataset

- 특징
  - https://github.com/NVlabs/ffhq-dataset
  - **StyleGAN3**에 사용
  - 1024*1024 크기의 70,000장의 png이미지
  - License: 상업적 이용만 불가

All data is hosted on Google Drive:

Path	Size	Files	Format	Description
ffhq-dataset	2.56 TB	210,014		Main folder
├ ffhq-dataset-v2.json	255 MB	1	JSON	Metadata including copyright info, URLs, etc.
├ images1024x1024	89.1 GB	70,000	PNG	Aligned and cropped images at 1024×1024
├ thumbnails128x128	1.95 GB	70,000	PNG	Thumbnails at 128×128
├ in-the-wild-images	955 GB	70,000	PNG	Original images from Flickr
├ tfrecords	273 GB	9	tfrecords	Multi-resolution data for StyleGAN and StyleGAN2
└ zips	1.28 TB	4	ZIP	Contents of each folder as a ZIP archive.

데이터셋 정보

7. ffhq-dataset

- 예시

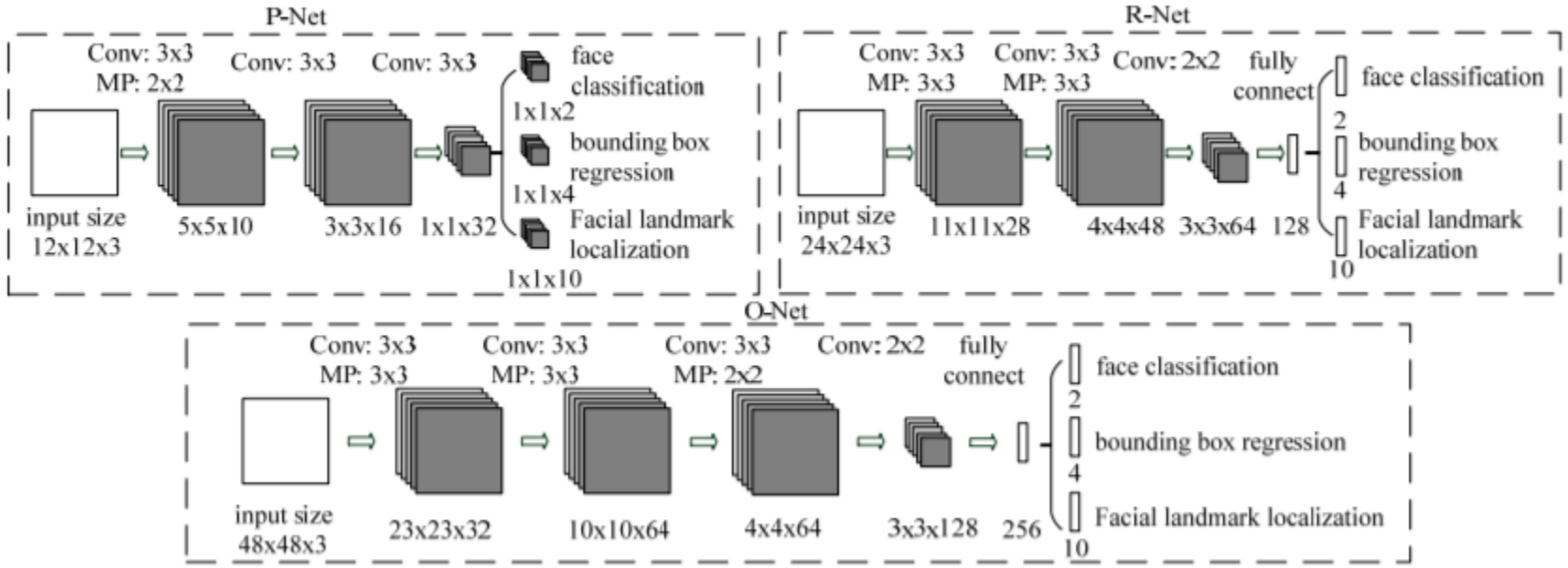
딥페이크 변조 영상 탐지 AI 경진대회 – 데이콘-

Face detection with MTCNN

Vector embedding with CNN (backbone : efficientnet b4)

Classifier with SVM

https://dacon.io/competitions/official/235655/codeshare/2351?page=1&dtype=recent



딥페이크 변조 영상 탐지 AI 경진대회 – 데이콘-

MTCNN 은 주로 P-Net, R-Net, O-Net의 3개로 분할된 CNN 모델

P-Net : 12*12 프레임에서 얼굴을 찾는다. 빠른 결과를 만드는 것

R-Net : 모든 후보가 전달되어 다량의 후보를 탈락시킨다

O-Net : 바운딩 박스와 얼굴 랜드마크 위치 반환

얼굴의 좌표 저장 (눈, 코, 입), 정확한 눈의 위치를 찾아 양쪽 눈 이 수평이 될 때까지 기본 이미지를 회전한다.

MTCNN은 얼굴 탐지 점수는 매우 높지만, 속도가 다른 모델에 비해 느리다. 속도 우선 시 SSD를 도입할 수 있다.

Deepfake Detection Challenge –캐글-

모델 Mesonet : 영상에서의 얼굴 조작을 탐지하기 위한

InceptionNet 기반 모델

얼굴 변조를 자동으로 효율적으로 감지하는 방법을 제시한다.

위조 동영상을 생성하는데 최근 두 가지 기술인 Deepfake 및 Face2Face에 중점을 둔다. 딥 러닝 접근 방식과 mesoscopic(중간) 속성에 초점을 맞추기 위해 적은 수의 레이어를 가진 두 개의 네트워크를 제시한다.

https://www.kaggle.com/fanbyprinciple/deep-fake-recognition-with-mesonet

Deepfake Detection Challenge –캐글-

모델 Facenet

얼굴 사진에서 그 사람에 대한 특징 값을 구해주는 모델로, 값들 간의 거리를 통해 이미지 identification, verification, clustering 할 수 있다.

- 모든 비디오에서 얼굴 추출
- 모든 이미지에서 embedding 벡터 계산
- PCA(주성분 분석)
- 밀도 기반 클러스터링
- 군집 분석 (비슷한 얼굴)
- 군집 준비 및 내보내기

https://www.kaggle.com/favadhassanjaskani/facenets-code

얼굴 생성 모델 SOTA – StyleGAN3

- GAN – 인위적인 샘플 합성

- StyleGAN – 매우 낮은 화질에서 시작해 높은 화질로, 인위적인 이미지를 점진적으로 생성 (이미지 생성 시 스타일 조절)

  Progressive Growing: 고해상도 이미지 생성하고

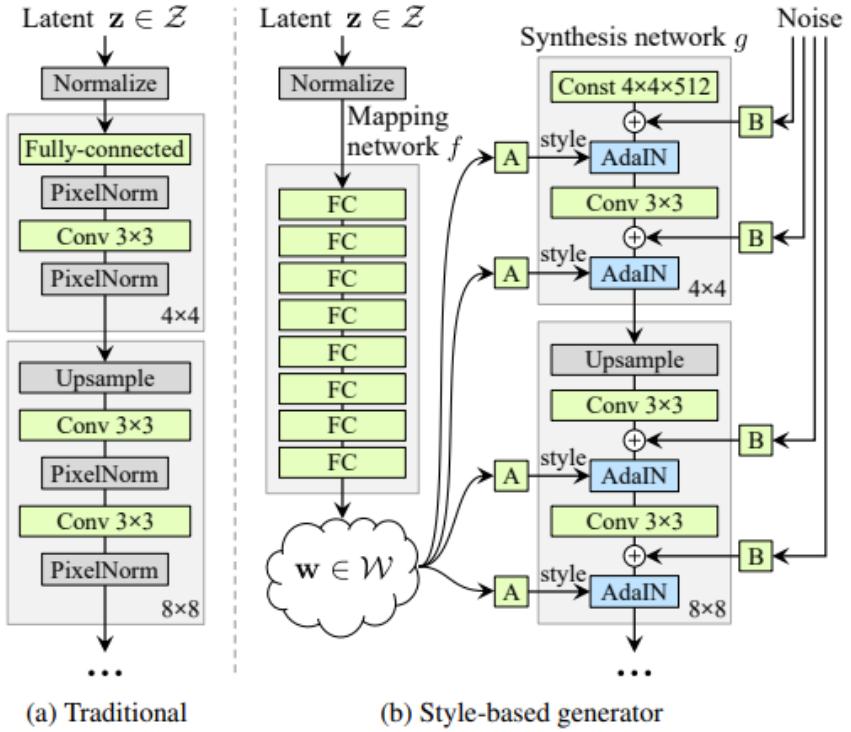
  AdaIn: 각 층에 이미지의 Style을 캡쳐하는 특징

  (스타일변환용 정규화 방법)

- StyleGAN2 – StyleGAN에서 AdaIn을 이용하면 feature map사이 크기 차이가 있을 경우, 정보를 왜곡한다. 따라서, AdaIn을 수정한 것

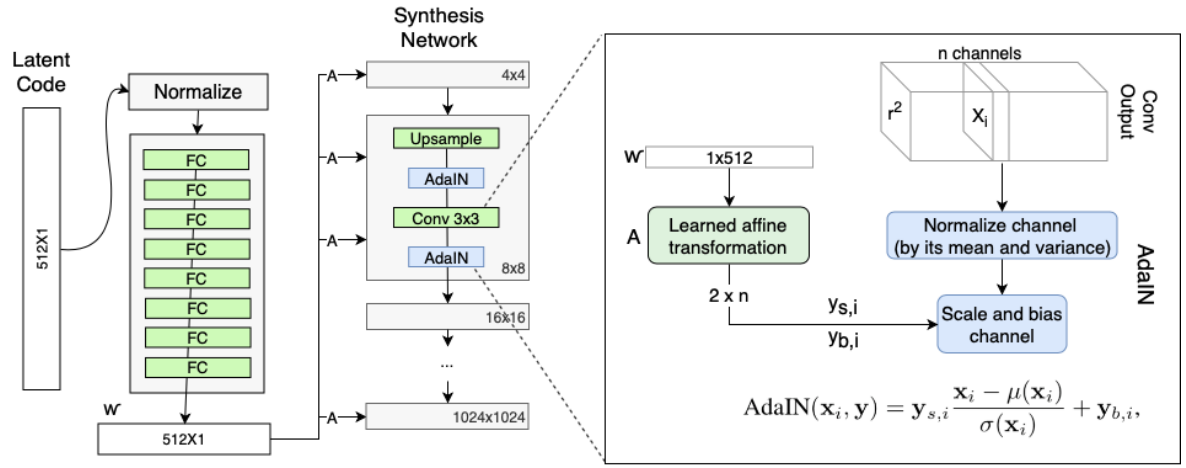
얼굴 생성 모델 SOTA – StyleGAN3

StyleGAN



(a) Traditional (b) Style-based generator

AdaIn



얼굴 생성 모델 SOTA – StyleGAN3

StyleGAN2에서 객체의 표면 대신 이미지 좌표에 접착된 것처럼 보이는 세부 사항 (머리카락, 주름 등)이 텍스처 세부 사항이 어떻게 고정되어 나타나지만

StyleGAN3는 Alias(대체,별명)가 없어 부드럽게 변환된다. Alias 없는 구조로 인해 신호는 이미지 기능의 모양과 상대적 위치를 모두 제어한다.

논문 제목 :

데이터 기반 딥페이크 탐지기법에 관한 최신 기술 동향 조사

- 게재된 저널 : **한국정보보호학회 (2020.10)**
- Abstract 정보 :

최근 전 세계적으로 '가짜뉴스', '가짜 연예인 음란 동영상' 및 '지인 능욕'에 사용되는 인공지능 기반의 딥페이크(Deepfakes)기술이 사회적인 이슈로 대두되고 있다. 딥페이크 기술이란 딥러닝 기술을 이용해 악의적으로 조작된 음성, 영상, 이미지 등을 만들어 내는 방법으로, 인공지능 기술의 발전에 맞추어 더욱더 빠르고 정교한 생성 기술이 등장하고 있다. 이러한 딥페이크 기술은 빠른 개발 속도와 쉬운 접근성을 기반으로 다양한 범죄에 악용되고 있다. 본 논문에서는 다양한 딥페이크 생성 기술을 설명하고, 이를 **효율적으로 탐지 할 수 있는 다양한 데이터 기반 딥페이크 탐지 기술의 현황을 설명**한다.

논문 제목 :

Gray 채널 분석을 사용한 딥페이크 탐지 성능 비교 연구

- 게재된 저널 : **한국멀티미디어학회 (2021.09)**
- Abstract 정보 :

최근 이미지 생성을 위한 딥러닝 기술의 개발로 정교한 딥페이크가 직접 생성되었다. 하지만 그 결과 딥페이크를 통한 사생활 침해도 증가했다. 이 문제를 해결하기 위해 주로 RGB 채널 기반 분석에 초점을 맞춘 딥페이크 탐지를 위한 여러 기술이 제안되었다. 기존 연구가 다른 색상 모델 기반 분석(즉, 그레이스케일)의 효과를 제안했지만, 그 효과는 아직 정량적으로 검증되지 않았다. 따라서 본 논문에서 우리는 **그레이스케일 채널 기반 분석의 효과를 딥페이크 검출의 RGB 채널 기반 분석과 비교**한다. 선택된 CNN 기반 모델과 딥페이크 데이터 세트를 기반으로 정확도와 시간 측면에서 각 색상 모델 기반 분석의 성능을 측정했다. 평가 결과는 그레이스케일 채널 기반 분석이 RGB 채널 분석보다 몇 가지 경우에 더 나은 성능을 발휘한다는 것을 확인했다.

논문 제목 :

Bidirectional Convolutional LSTM을 이용한 Deepfake 탐지 방법

- 게재된 저널 : **한국정보보호학회 (2020.12)**
- Abstract 정보 :

최근 하드웨어의 성능과 인공지능 기술이 발달함에 따라 육안으로 구분하기 어려운 정교한 가짜 동영상들이 증가하고 있다. 인공지능을 이용한 얼굴 합성 기술을 딥페이크라고 하며 약간의 프로그래밍 능력과 딥러닝 지식만 있다면 누구든지 딥페이크를 이용하여 정교한 가짜 동영상을 제작할 수 있다. 이에 무분별한 가짜 동영상이 크게 증가하였으며 이는 개인 정보 침해, 가짜 뉴스, 사기 등에 문제로 이어질 수 있다. 따라서 사람의 눈으로도 진위를 가릴 수 없는 가짜 동영상을 탐지할 수 있는 방안이 필요하다. 이에 본 논문에서는 **Bidirectional Convolutional LSTM과 어텐션 모듈(Attention module)을 적용한 딥페이크 탐지 모델을 제안**한다. 본 논문에서 제안하는 모델은 어텐션 모듈과 신경곱 합성망 모델을 같이 사용되어 각 프레임의 특징을 추출하고 기존의 제안되어왔던 시간의 순방향만을 고려하는 LSTM과 달리 시간의 역방향도 고려하여 학습한다. 어텐션 모듈은 합성곱 신경망 모델과 같이 사용되어 각 프레임의 특징 추출에 이용한다. 실험을 통해 본 논문에서 제안하는 모델은 93.5%의 정확도를 갖고 기존 연구의 결과보다 AUC가 최대 50% 가량 높음을 보였다.

논문 제목 :

Deepfakes Detection With Automatic Face Weighting

- 게재된 저널 : **IEEE/CVF**
- Abstract 정보 :

변경되고 조작된 멀티미디어는 점점 더 존재하고 소셜 미디어 플랫폼을 통해 널리 배포된다. 고급 비디오 조작 도구를 사용하면 매우 사실적으로 보이는 변경된 멀티미디어를 생성할 수 있다. 조작을 감지하기 위해 많은 방법이 제시되었지만 대부분은 연구 환경에서 사용되는 데이터 세트 외부의 데이터로 평가할 때 실패한다. 이 문제를 해결하기 위해 DFDC(Deepfake Detection Challenge)는 실제 조작이 포함된 대규모 비디오 데이터 세트와 까다로운 데이터에 직면했을 때도 방법이 빠르고 정확하게 작동하도록 보장하는 평가 시스템을 제공한다. 이 논문에서는 우리는 조작을 정확하게 감지하기 위해 **비디오에 존재하는 얼굴에서 시각적 및 시간적 특징을 추출하는 CNN(Convolutional Neural Networks) 및 RNN(Recurrent Neural Networks)을 기반으로 하는 방법을 소개**한다. 이 방법은 DFDC 데이터 세트로 평가되어 다른 기술에 비해 경쟁력 있는 결과를 제공한다.

논문 제목 :

Deepfakes Detection Through Precise Geometric Features

- 게재된 저널 : **IEEE/CVF**
- Abstract 정보 :

딥페이크(Deepfake)는 비디오의 대상 얼굴을 원본에 이식하는 악성 기술의 한 분야로 저작권 침해, 정보의 혼란, 심지어 공황상태와 같은 심각한 문제를 초래한다. 딥페이크 비디오 감지를 위한 이전의 노력은 정교한 조작에 의해 우회될 위험이 있는 외관 특징에 주로 집중되어 모델 복잡성과 노이즈에 대한 민감도가 높다. 게다가, 조작된 비디오의 시간적 특징을 어떻게 채굴하고 그것들을 어떻게 이용할 것인가는 여전히 미해결 문제이다. 우리는 **정확한 기하학적 특징에 대한 시간적 모델링을 통해 딥페이크 비디오를 탐지하기 위한 LRNet이라는 효율적이고 강력한 프레임워크를 제안**한다. 새로운 보정 모듈은 기하학적 형상의 정밀도를 향상시켜 차별성을 높이고, 시간적 형상의 충분한 활용을 위해 2 스트림 반복 신경망(RNN)을 구성한다. 이전 방법에 비해 우리가 제안한 방법은 무게가 가볍고 훈련하기 쉽다. 또한, 우리의 방법은 고도로 압축되거나 소음이 손상된 비디오를 감지하는 데 견고함을 보여주었다. 우리 모델은 FaceForensics++ 데이터 세트에서 0.999 AUC를 달성했다. 한편, 고압축 영상에서는 성능 저하(-0.042AUC)가 완만하다.

논문 제목 :

Detecting Deepfakes with Metric Learning

- 게재된 저널 : **IEEE**
- Abstract 정보 :

FaceApp, SnapChat, MixBoot, FaceBlender 등과 같은 여러 페이스 스왑 애플리케이션이 등장함에 따라 디지털 미디어 콘텐츠의 진정성은 매우 느슨한 실타래에 걸려 있다. 소셜 미디어 플랫폼에서는 동영상이 높은 압축률로 널리 유통된다. 본 연구에서는 **고압축 시나리오에서 딥페이크 분류의 맥락에서 몇 가지 딥러닝 접근 방식을 분석하고 메트릭 학습을 기반으로 제안된 접근 방식이 그러한 분류를 수행하는 데 매우 효과적일 수 있음을 입증**한다. 사실성을 평가하기 위해 비디오당 프레임 수를 줄임으로써, 트리플렛 네트워크 아키텍처를 사용한 메트릭 학습 접근 방식은 성과가 있는 것으로 입증되었다. 벡터를 내장한 실제 비디오와 가짜 비디오 클러스터 사이의 피처 공간 거리를 향상시키는 방법을 학습한다. 서로 다른 환경의 동작을 분석하기 위해 두 데이터 세트에 대한 접근 방식을 검증했다. 우리는 Celeb-DF 데이터 세트에서 99.2%, 고도로 압축된 신경 텍스처 데이터 세트에서 90.71%의 정확도를 달성했다. 우리의 접근 방식은 데이터 압축이 불가피한 소셜 미디어 플랫폼에서 특히 유용하다.

논문 제목 :

MTD-Net: Learning to Detect Deepfakes Images by Multi-Scale Texture Difference

- 게재된 저널 : **IEEE**
- Abstract 정보 :

안면조작 기술이 빠르게 발전하면서 사람의 눈으로는 가짜 얼굴상을 구별하기 어렵다. 반대로 CNN(Convolutional Neural Network) 판별기는 가짜/실제 얼굴 이미지를 식별하는 데 높은 정확도에 빠르게 도달할 수 있다. 이 연구에서는 **가짜/실제 얼굴을 구별하는 CNN 모델의 행동을 탐구**한다. 우리는 **다중 스케일 텍스처 차이 정보가 얼굴 위변조 탐지에 중요한 역할을 한다는 것을 발견**했다. 위의 관찰에 힘입어, **우리는 강력한 얼굴 위조 탐지를 위해 MTD-Net으로 만들어진 새로운 다중 스케일 텍스처 차이 모델을 제안**한다. 이 모델은 중앙 차이 컨볼루션(CDC)과 애터러스 공간 피라미드 풀링(ASPP)을 활용한다. CDC는 픽셀 강도 정보와 픽셀 그라데이션 정보를 결합하여 텍스처 차이 정보에 대한 고정된 설명을 제공한다. 동시에 ASPP를 기반으로 하는 다중 스케일 정보 융합은 텍스처 기능이 파괴되는 것을 막을 수 있다. FaceForensics++, DeepForensics-1.0, Celleb-DF 및 DFDC와 같은 여러 데이터베이스에 대한 실험 결과는 MTD-Net이 기존 접근 방식을 능가한다는 것을 증명한다. MTD-Net은 JPEG 압축 및 블러와 같은 이미지 왜곡에 더 강력하며, 이는 긴급히 필요하다.

논문 제목 :

A Deepfakes detection technique based on two-stream network

- 게재된 저널 : **IEEE**
- Abstract 정보 :
- 딥러닝 기술이 빠르게 발전하면서 딥페이크와 같은 딥 위작 기술이 인터넷 구석구석을 가득 메우기 시작했다. Deepfake는 생성적 적대 네트워크와 자동 인코더 기술을 활용하여 얼굴을 대체하고 표정을 쉽게 변조한다. 딥페이크는 가짜 음란물을 제작하고, 소문을 퍼뜨리고, 가짜 뉴스를 퍼뜨리고, 심지어 정치 선거에까지 영향을 미쳐 참담한 사회적 결과를 초래할 수 있다. 다만 이런 종류의 가짜 동영상에 대한 탐지 기술은 아직 세대 기술에 한참 뒤처져 있고, 기존 작품들은 다소 한계가 있다. 본 논문은 먼저 **기존 생성 및 탐지 작업을 요약하고 기존 작업의 결함을 분석한 다음 EfficientNet을 기반으로 하는 2스트림 네트워크 탐지 프레임워크를 제안**한다. 대규모 오픈 소스 데이터 세트인 FaceForensics++에서 테스트함으로써 탐지 방법은 평균 99% 이상의 정확도로 가짜 비디오를 탐지하고 압축에 대한 모델의 기능을 어느 정도 개선할 수 있었다.