캡스톤 디자인
‘딥페이크 탐지’

#4. 딥페이크 관련 코드 조사

김지수, 김민지, 민지민

지난 캡스톤 회의 내용 - 딥페이크 탐지 우회 기술 논문 조사-

- Adversarial Attack (적대적 공격)

DeepFake 탐지기에 대한 적대적 공격을 수행

출력 다양화 초기화에 기초한 CycleGAN 적대적 공격

- Adversarial Attack 방어

MagDR (마스크 유동 탐지 및 재구성) 을 이용해 딥페이크의 얼굴 교체, 편집, 재현을 방어

변환 인식 적대적 얼굴을 통한 GAN 기반 Deepfake 공격 방어

지난 캡스톤 회의 내용 - 딥페이크 탐지 우회 기술 논문 조사-

- **Poisson Noise DeepFool**을 사용해서 최소한의 perturbation으로 예측을 변화시킴으로써 딥페이크 탐지기 정확도 감소

- **Implicit Spatial-Domain Notch Filtering**을 통해 이미지 품질을 손상시키지 않고 fake 이미지의 artifact 패턴을 줄여 기존 탐지기의 정확도 감소

- **White-Box, Black-Box** 공격

지난 캡스톤 회의 내용 - 딥페이크 탐지 우회 기술 논문 조사-

- **FakePolishier**를 사용하여 아티팩트를 줄여 탐지 정확성 감소
- 가짜 비디오를 적대적으로 수정하여 탐지기 우회
- 얼굴 매니폴드에서 적대적 포인트를 최적으로 검색하여 안티포렌식 가짜 이미지 생성

⇒ **이번주 목표**

- 딥페이크 생성 코드 찾아보고 실행해보기
- 안티 포렌식 생성 코드 찾아보고 실행해보기

⇒ **실제 수행한 내용**

- 딥페이크 생성 코드 찾아보고 실행해보기 (O)
- 안티 포렌식 생성 코드 찾아보고 실행해보기 (x) → 코드 거의 없었음

⇒ **문제점**

- 텐서플로우나 쿠다 버젼과 같은 환경 문제에 봉착하여 실행할 수 있는 코드 많지 않았음
- gpu가 없어서 코랩 관련 코드만 실행해봄

코랩을 이용한 딥페이크 생성

github에서 clone 하여 사용

이미지와 영상을 256*256 해상도로 설정, 영상은 30초 이내

영상제작에 필요한 사진, 동영상 및 인공지능 신경망 파일을 구글 드라이브에 연결하여

입력 사진과 입력 영상을 불러온 후 영상을 만듭니다.

vox-cpk.pth.tar 라는 인공지능 신경망 이용 → 이는 딥페이크에 필요한 모델 파일들을 모아놓은 아카이브이다.

https://github.com/drminix/first-order-model

코랩을 이용한 딥페이크 생성

또한, 사진에서 얼굴을 추출해서 딥페이크 영상을 만들 수 있습니다.

다음과 같은 사진에서 얼굴을 추출하여 같은 방식으로 만들어보았다.

DeepFaceLab을 이용한 딥페이크 생성

https://github.com/iperov/DeepFaceLab

딥페이크 프로그램인 딥페이스랩을 이용하여 딥페이크 영상을 생성

현재 github에 공개된 오픈소스로 배치파일들을 차례대로 실행하여 딥페이크 영상 생성

DeepFaceLab을 이용한 딥페이크 생성

소스영상

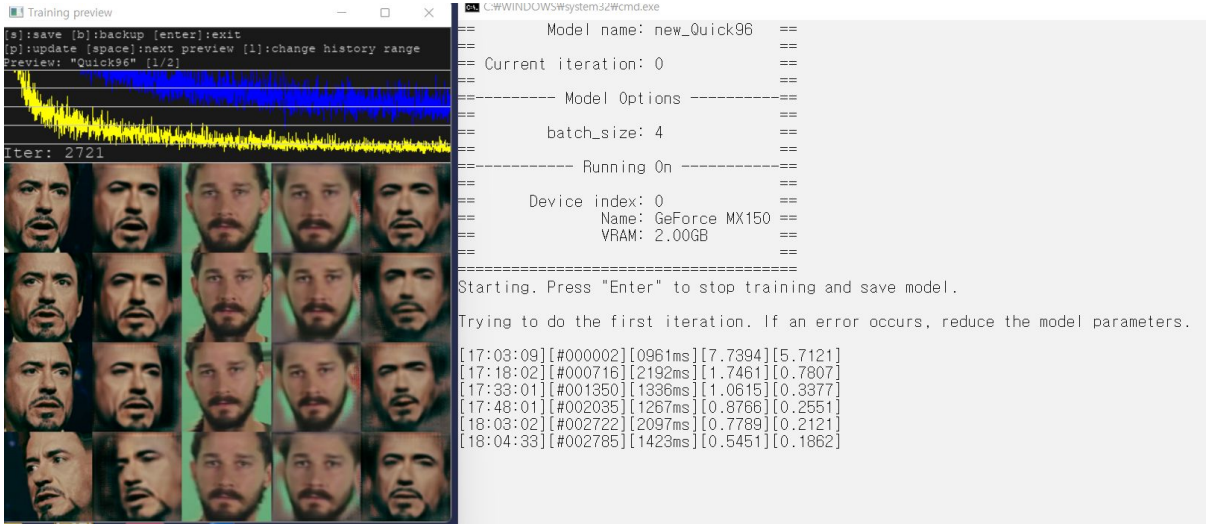
목적영상

→ 목적영상에 소스영상의 얼굴을 합성

→ 먼저, 각 영상에서 프레임 이미지 추출, 얼굴 이미지 추출하는 작업 진행

DeepFaceLab을 이용한 딥페이크 생성

트레이닝 단계



소스영상얼굴과 목적영상 얼굴의 얼굴을 학습하는 딥러닝 과정

→ 트레이닝 시간이 길수록 좀 더 선명해지고 완성도 있는 완성도있는 얼굴표정이 만들어지게 됨

DeepFaceLab을 이용한 딥페이크 생성

목적 영상에 소스 영상의 얼굴을 합성한 결과

SimSwap을 이용한 딥페이크 face swap 이미지 생성

https://github.com/neuralchen/SimSwap

- ResNet 기반 알고리즘

- image-image → image
- image-video → video

- Anaconda Prompt 기반 실행 시도

  → 실패(정확한 이유X, 아마 환경 문제)

- Google Colaboratory 기반 코드 발견

SimSwap을 이용한 딥페이크 face swap 이미지 생성

실행 환경: Google Colaboratory

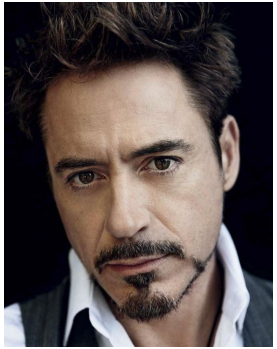
- GitHub 코드 clone하여 실행
- GitHub 안에 실행 가능한 데이터셋 존재
- 외부 데이터 연동해서 실행도 가능

<image-image → image>

→

SimSwap을 이용한 딥페이크 face swap 이미지 생성

<image-video → video>



→

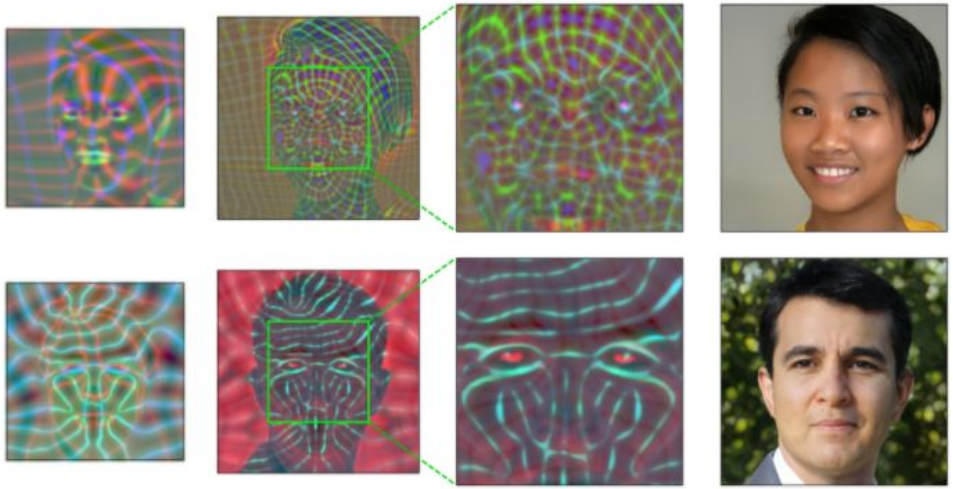
딥페이크 generation 개발물

FaderNetworks (facebookresearch 2019) - https://github.com/facebookresearch/FaderNetworks

딥페이크 generation 개발물

StyleGAN3(NVlabs) - https://nvlabs.github.io/stylegan3/

- StyleGan을 통해서 고해상도의 가짜이미지 생성
- 이전에는 불가능했던 style의 scale-specific control 가능



AIHub의 딥페이크 변조 영상 데이터셋 생성 알고리즘

모델명	설명	수량
DeepFaceLab (이하 DFL)	가장 대중적인 모델이며, 평균적인 완성도가 가장 높은 Face Swapping 모델.	53,816개 생성 (목표 수량 37,500개)
DeepFakes/FaceSwap (이하 DFFS)	최초의 딥페이크 모델들 중 하나이자 결과물이 안정적인 Face Swapping 모델	52,209개 생성 (목표 수량 37,500개)
Face Swapping GAN (이하 FSGAN)	비교적 최근에 공개된 생성적 적대 신경망 기반 Face Swapping / Reenactment 모델	53,816개 생성 (목표 수량 37,500개)
First Order Model (이하 FO)	영상 움직임 키포인트 기반 Face Reenactment 모델	72,298개 생성 (목표 수량 37,500개)
Audio-driven Talking Face Video Generation with Learning-based Personalized Head Pose (이하 Audio-Driven)	비교적 최근에 공개된 음성입력 및 3D 모델링을 통해 얼굴영상을 생성하는 Face Reenactment 모델 (음성을 사용하는 공통점 때문에 아래 모델과 목표수량을 공유하도록 함)	21,731개 생성 (목표 수량 3,000개)
Wav2Lip (이하 Audio-driven)	가장 최근에 공개된 음성입력 기반 Face Reenactment 모델	

시간 상 직접 실행X