FakeME: 한국인 기반 DeepFake 생성 및 탐지 시스템

2024-2 AJOU SOFTCON

팀 명 FakeME **팀 원** 박주연,곽세현 **지도교수** 박정훈 교수 **멘 토** 이상록 멘토님

개발 동기 및 목적

[DeepFake란?]

Deep Learning과 Fake의 합성어, 딥러닝 기술을 사용하는 인간 이미지 합성 기술
생성적 적대 신경망 GAN 기술 발달로 인해 DeepFake의 퀄리티가 상승

딥페이크 범죄 발생 건수,검거 현황
(자료: 경찰청, 더불어민주당 황정아 의원)


넷플릭스 오리지널 시리즈 '살인자ㅇ난감'에서 배우 손석구와
딥페이크 기술로 만든 어린 시절의 손석구

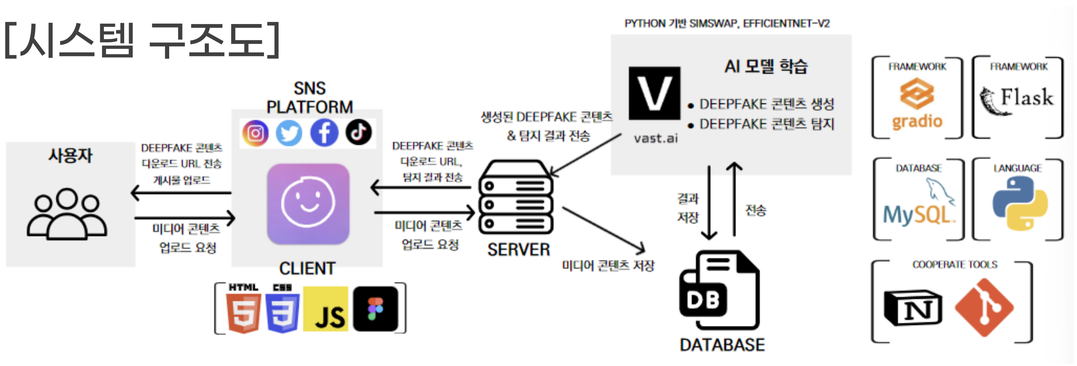

앳되고 힘찬 유관순 열사 목소리…
AI로 되살아난 독립운동가

디지털 성범죄, 허위 정보 유포 등으로 인해 DeepFake 기술에 대한 부정적 인식 확산하지만, Deepfake는 광고, 영화 등의 분야에서 창의적으로 활용 가능하며 비용 절감 등 콘텐츠 제작의 효율을 높이는 긍정적 역할 수행

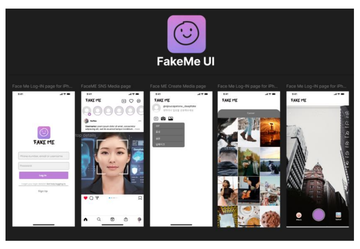
→ Deepfake의 긍정적인 잠재력을 극대화, 부정적인 영향을 최소화하여 기술의 양면성 해결

개발 내용

[시스템 구조도]



[시스템 사용자 인터페이스 (UI) - 모바일 기반]



- SNS 플랫폼 내에서 시스템을 제공
- 단일 플랫폼 내에서 Deepfake 생성, 탐지 기능 동시 제공으로 사용자에게 편의 제공
- 기존 SNS 플랫폼과 위화감이 없는 UI로 사용자 친화적인 서비스 제공

[DeepFake 생성]

- 유저가 미디어 콘텐츠를 업로드하면 생성 모델을 통해 Deepfake 콘텐츠 생성
- 5초 내로 빠른 Deepfake 콘텐츠 생성
- JPG형태로 다운로드 URL 제공

[DeepFake 탐지]

- 유저가 게시물을 업로드하면 탐지 모델을 통해 Deepfake 콘텐츠 여부 탐지
- 1초 내로 빠른 Deepfake 콘텐츠 탐지
- 해당 콘텐츠가 Deepfake일 시 #DeepFake 태그 추가 후 게시물 업로드

주요기술

[데이터 전처리]


AI허브 딥페이크 변조
영상 데이터


Kaggle – FaceForensics++
VGG Face2 – 고화질 Data


AI허브 – 한국인 안면 이미지

1. Video에서 image 변환 후 mtcnn , facenet을 활용한 얼굴 인식
2. 224 x 224 size로 crop
3. Rotation, Flip 등 Data augmentation & Image Normalization

[SimSwap 기반 Deepfake 생성]


기존의 SimSwap
(디카프리오+백종원)


튜닝후의 SimSwap
(디카프리오+백종원)


튜닝후의 SimSwap
비가시적 워터마크 삽입

(1) 추가 튜닝 진행
- 정확도와 성능을 개선하기 위해 특정 레이어에 필요한 파라미터 추가
- 모델 학습 안정성을 위해 초기화되지 않은 레이어에 대해 weight_init 함수를 적용하여 안정적인 가중치 초기화 수행
- Deepfake 품질 개선을 위한 파라미터 및 초기화 설정 세부 조정
- OpenCV를 통한 Skin Smoothing 진행
- 전이학습을 통한 생성모델 한국인 최적화

(2) 비가시적 워터마크 삽입
- 탐지 보조를 위해 생성시 비가시적 워터마크 삽입
- 텍스트를 기반으로 템플릿 이미지 생성, 이미지 내에 랜덤하게 배치
- GRB중 R의 값이 150 이상일 경우 해당 픽셀을 검은색으로 설정, 이외의 경우는 흰색으로 설정 → 이를 기반으로 워터마크 복원

[EfficientNet-V2 기반 Deepfake 탐지]

	Test Accuracy	Precision FAKE	Precision REAL	Recall FAKE	Recall REAL	F1-Score FAKE	F1-Score REAL
Our model (EfficientNet-V2)	94.78%	0.98	0.92	0.92	0.98	0.95	0.95
VGG16	87.58%	0.90	0.86	0.85	0.90	0.87	0.88
XceptionNet	90.56%	0.91	0.90	0.90	0.91	0.91	0.91

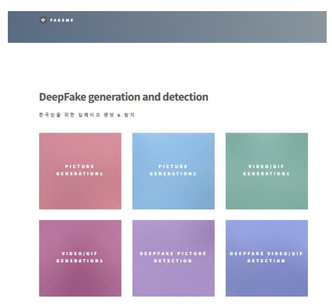
(1) Backbone 모델 구조
- Conv 3x3을 통한 입력 이미지에서 저수준 특징(가장자리, 텍스처 등) 추출
- MBConv 를 활용하여 Deepfake의 미세한 특성을 학습
- Fused-MBConv를 활용하여 공간적 및 채널적 정보의 통합 학습

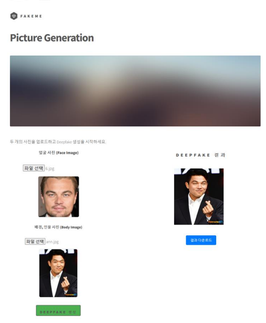
(2) Fine-Tuning 진행
- Learning rate, Optimizer, Dropout 등 학습 하이퍼파라미터 조정
- 사전 학습 모델을 통해 한국인 데이터에 적합하게 전이 학습 진행

(3) 데이터 구성
- Train Dataset : Test Dataset = 24,000 : 6,000으로 구성

결과 및 분석

[Deepfake 생성]



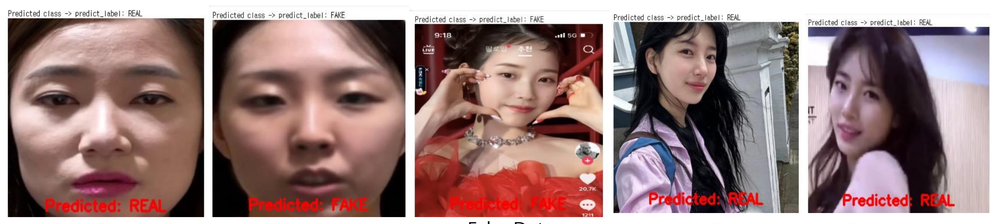



원본 image
동양인(김수현) DeepFake image
서양인(디카프리오) DeepFake image


원본 image
DeepFake적용 워터마크 X
DeepFake적용 워터마크 O

[Deepfake 탐지]

Class	Precision	Recall	F1-Score	Support
Fake	0.98	0.92	0.95	3000
Real	0.92	0.98	0.95	3000



Real Data
Predicted Real
Fake Data
Predicted Fake
Fake Data
Predicted Fake
Real Data
Predicted Real
Real Data
Predicted Real

오픈소스 URL

https://github.com/AjouCapstoneforDeepFake/DeepFake