딥페이크 탐지모델개발

캡 스 톤 2 계 획 발 표

20191769 김지수, 20191767 김민지, 20191730 민지민

CONTENTS

01

캡스톤 디자인2
배경 및 필요성

01. 딥페이크
02. 현황
03. 문제점
04. 목표 및 기대효과

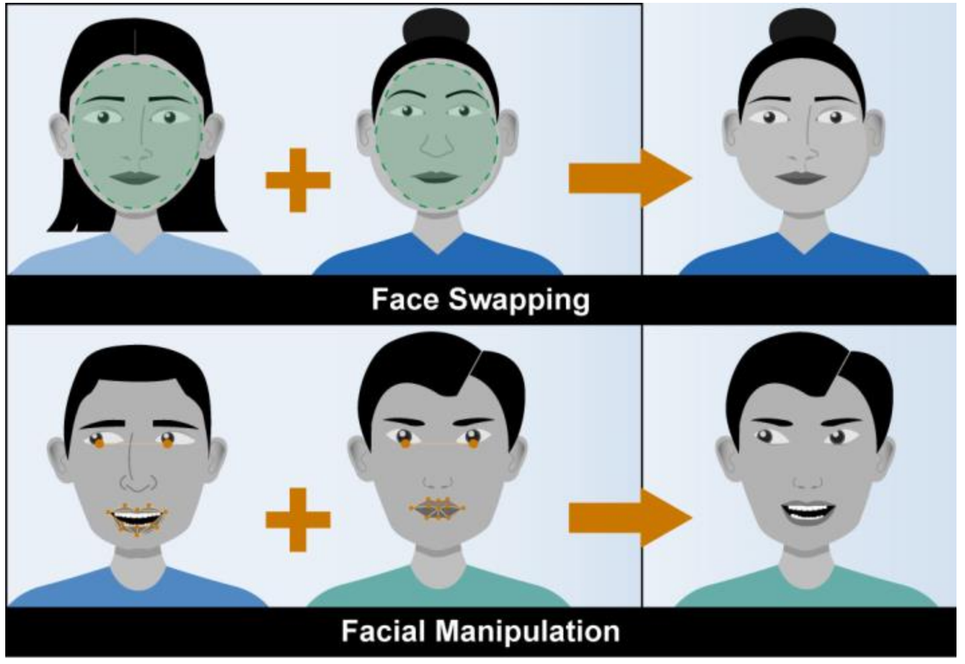
01 배경 및 필요성

딥페이크

딥페이크는 딥러닝과 페이크의 합성어

적대관계생성신경망(GAN: Generative Adversarial Network)이라는 기계학습(ML) 기술을 사용하여,

기존 사진이나 영상을 원본에 겹쳐서 만듦



01 배경 및 필요성

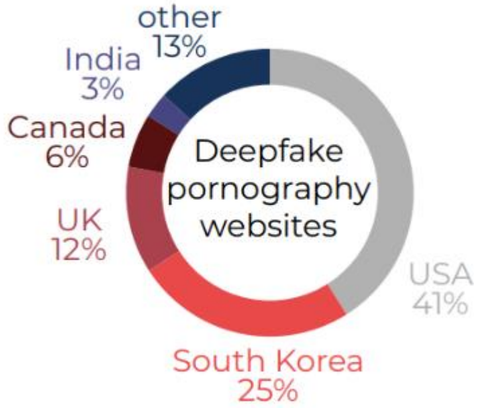
딥페이크의 현황

2019년 네덜란드의 보안업체 Deeptrace의

" The state of deepfakes 2019 "

온라인 딥페이크 사용 목적 비율

96% 포르노그래피 4% 교육 및 기타



피해자 중 25%가 한국 여성이며

연예인과 일반인 모두 해당

01 배경 및 필요성

딥페이크의 현황

톰 크루즈가 아이언맨이 된 영상

딥페이크 기술 → '개인 정보 침해', '사기' 등 **다양한 범죄에 악용**

그에 따라, 딥페이크 영상 탐지에 대한 필요성이 대두

⇒ **영상 변조 · 합성 탐지**에 대한 연구가 활발히 진행되고 있음

01 배경 및 필요성

문제점

딥페이크 탐지기가 **간단한 이미지 변형만으로도 탐지가 우회된다는 문제점**

⇒ 딥페이크 우회 기술 탐지에 대한 연구는 현저히 적음

⇒ **탐지가 어려운 딥페이크의 생성 및 확산을 방지하기 위한** 우회 기법에 대한 고찰과 그렇게 생성된 이미지에 대한 탐지율 향상을 목표로 한 연구가 필요

01 배경 및 필요성

안티포렌식

- 탐지 기술 무력화 → **안티포렌식**
- 적대적 데이터셋을 생성하는 안티 포렌식 공격

화이트 박스 공격(White - Box Attacks)

- 공격자가 탐지 모델에 대한 정보를 모두 안다는 전제
- 비현실적 조건
- 공격 성공률 100% 가까움

블랙 박스 공격(Black - Box Attacks)

- 공격자가 원본 이미지에 특정 노이즈 추가 → 오분류 유도
- 실제 화이트 박스 공격보다 많이 시도됨

01 배경 및 필요성

두가지의 블랙 박스 공격

- **Adversarial Attack**
  - 데이터셋에 네트워크를 교란시키는 노이즈를 추가
  - 초보자가 쉽게 공격할 수 있는 방법 X

- **이미지 편집 기법**
  - 데이터셋에 이미지 편집을 가함
  - 초보자가 쉽게 공격할 수 있는 방법 O

⇒ 두 방법 모두
딥페이크 탐지기 무력화 가능

⇒ 전문가가 아니더라도 공격할 수 있는
이미지 편집 기법 기반의 블랙 박스 공격에 대응할 기술 필요

01 배경 및 필요성

01. 딥페이크

02. 현황

03. 문제점

04. 목표 및 기대효과

캡스톤 디자인 목표 및 기대효과

' 여러 탐지 우회기술인 안티포렌식 기법에도 강인한
딥페이크 탐지 모델 생성 '

안티포렌식 기법으로 조작된 이미지나 영상을 좀 더 민감하게 판별할 수 있고,
더 나아가 관련 범죄 발생률을 줄일 수 있음

01 배경 및 필요성

01. 딥페이크

02. 현황

03. 문제점

04. 목표 및 기대효과

캡스톤 디자인 목표 및 기대효과

' 여러 탐지 우회기술인 안티포렌식 기법에도 강인한 딥페이크 탐지 모델 생성 '

캡스톤 디자인 1 : 안티포렌식 공격기법을 **학습에 반영하는** 방법으로 강인한 탐지모델 생성

캡스톤 디자인 2 : **학습에 반영하지 않은** 안티포렌식 공격기법에 대해서도 강인한 모델 생성

02

캡스톤 디자인2
내용

01. 주요기능
02. 학습절차

02 내용

주요 기능

다양한 유형의 안티포렌식 데이터셋에 대한 높은 딥페이크 탐지율을 도출

⇒ 다양한 안티포렌식 공격기법의 공격 종류와 공격 강도를 랜덤하게 적용한 데이터셋을 생성 후 적대적 학습에 사용

학습에 반영한 공격 + 학습에 반영하지 않은 공격

⇒ 모두에 강인한 모델 생성

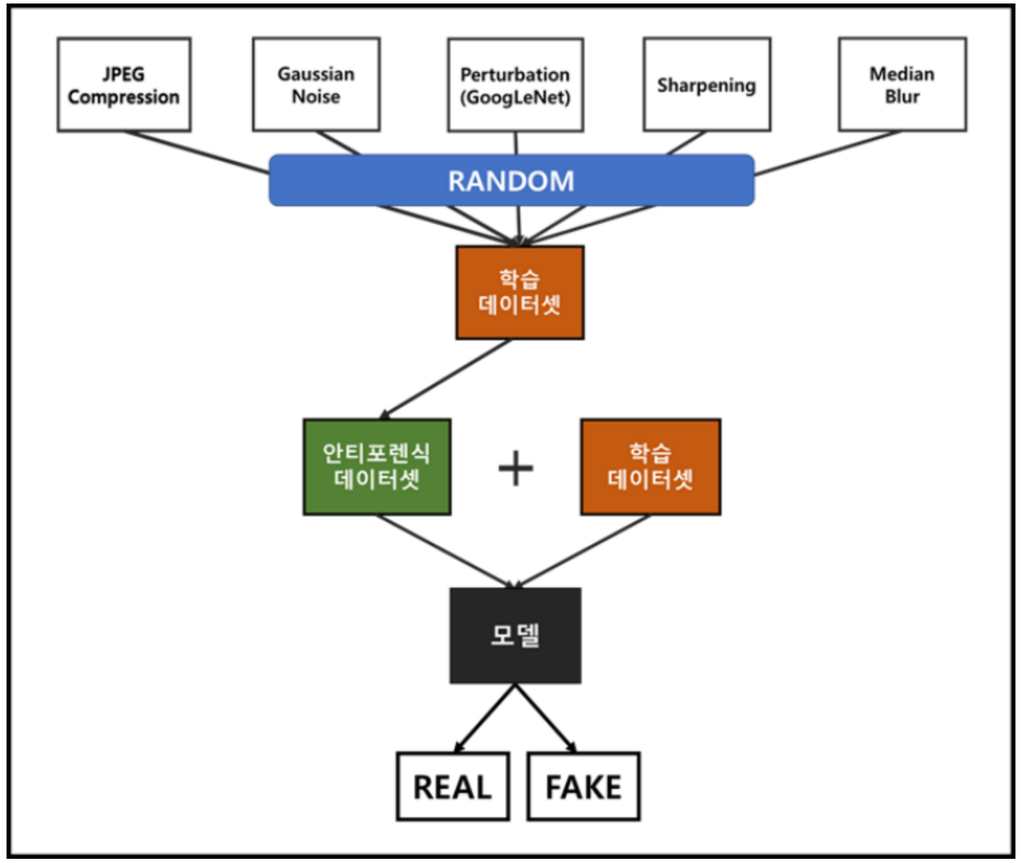
02 내용

01. 주요 기능

02. 학습절차

학습절차

- 제안하는 적대적 학습 기법 도면



03

캡스톤 디자인2
추진 전략

01. 예상 문제점
02. 진행사항
03. 역할 분담
04. 일정

03 추진전략

예상 문제점

컴퓨터비전 프로젝트에서 자주 대두되는 문제 : **데이터셋의 저작권 문제**

⇒ AI허브의 개방된 학습용 데이터셋을 사용

03 추진전략

진행사항

- 다양한 안티포렌식 기법이 적용된 데이터 셋 생성

Sharpening
Median blur
GoogLeNet Perturbation
Gaussian Noise
JPEG Compression

→ 공격 순서 / 공격 강도 / Random →

- 적대적 학습 전 딥페이크 탐지 성능

Validation loss	Validation Accuracy
0.0015	1.000

03 추진전략

진행사항

- **앞으로의 계획 :**
  학습도면 절차에 따라 안티포렌식 데이터셋을 적대적 학습에 적용하여 **학습에 반영한 안티포렌식 공격 뿐만 아니라 학습에 반영하지 않은 공격들에 대해서도 강인한 모델 생성**

03 추진전략

역할 분담

- 각자 주어진 환경에 맞추어 업무 부담

김지수
(팀장)

실험 데이터셋 생성

김민지

학습 모델 생성

민지민

학습 모델로 성능 확인

03 추진전략

일정

월	6		7				8					9				10				11	
주	4	5	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	3	4	1	2
계획서 작성	■																				
계획발표		■																			
안티 포렌식 데이터셋 생성			■	■	■	■															
실험진행							■	■	■	■	■	■	■	■	■	■	■	■			
중간 보고서 작성										■											
중간발표											■										
최종발표																			■	■	■

감사합니다

20191769 김지수, 20191767 김민지, 20191730 민지민