딥페이크
캡 스 톤 I 계 획 발 표
탐지모델개발

20191769 김지수, 20191767 김민지, 20191730 민지민

CONTENTS

01

주요
키워드 설명

01. 딥페이크
02. 디지털 포렌식
03. 안티 포렌식

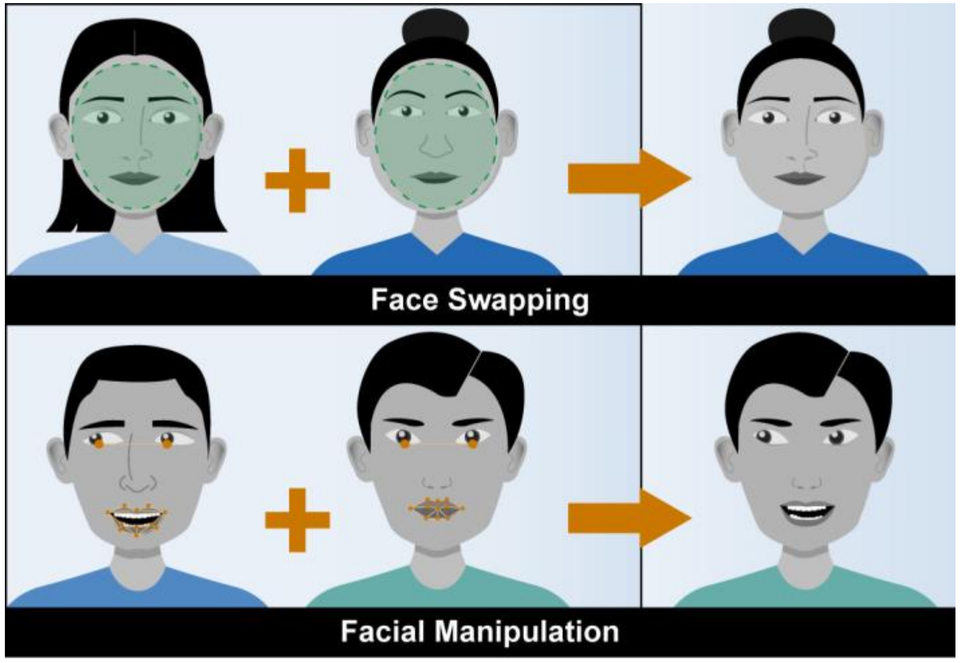
01 주요 키워드 설명

딥페이크

딥페이크는 딥러닝과 페이크의 합성어

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

기존 사진이나 영상을 원본에 겹쳐서 만듦



01 주요 키워드 설명

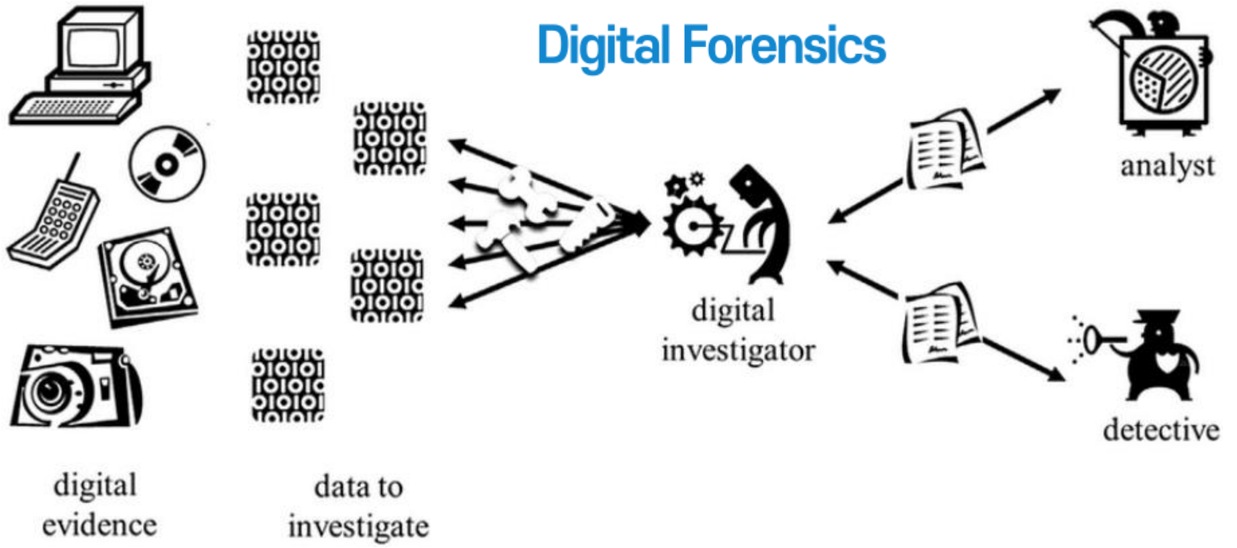
01. 딥페이크

02 . 디지털 포렌식

03. 안티 포렌식

디지털 포렌식 및 안티 포렌식

디지털포렌식은 PC나 휴대폰 등에 남아 있는 디지털 정보를 분석해 범죄 단서를 찾는 수사기법, 반대로 **안티포렌식**은 디지털 정보를 분석하지 못하도록 삭제하는 기술



02

캡스톤 디자인
배경 및 필요성

01. 현황
02. 문제점
03. 목표

02 배경 및 필요성

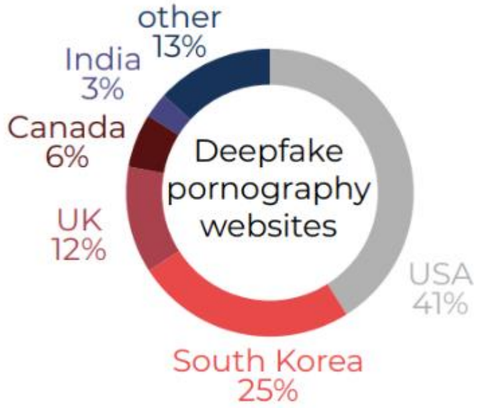
딥페이크의 현황

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

02 배경 및 필요성

01 . 현황

02. 문제점

03. 목표

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

02 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

02 배경 및 필요성

01. 현황

02. 문제점

03. 목표

캡스톤 디자인 목표

' 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

03

캡스톤 디자인
내용

01. 주요기능
02. 블랙박스 기법
03. 화이트박스 기법

03 내용

주요 기능

다양한 안티포렌식 기법으로 생성된 데이터셋에 대한 높은 딥페이크 탐지율을 도출

⇒ 원본 데이터와 직접 생성한 안티포렌식이 적용된 데이터셋을 함께 이용해서 adversarial training 후 fine-tuning을 진행

03 내용

01. 주요 기능

02. 블랙박스 기법

03. 화이트박스 기법

블랙박스 기법

블랙-박스 기법은 공격자가 모델에 대한 정보를 알지 못하고 공격하는 것

ex) denoise, jpeg compression, sharpening, gaussian noise, salt and pepper noise

원본 이미지

gaussian noise
(gauss_var = 1000)

03 내용

01. 주요 기능

02. 블랙박스 기법

03. 화이트박스 기법

블랙박스 기법

원본 이미지

sharpening
(sharpening_arr = 9)

원본 이미지

salt and pepper noise
(p = 0.005)

03 내용

블랙박스 기법

원본 이미지

denoise
(h = 10)

원본 이미지

jpeg compression
(quality = 10)

03 내용

화이트박스 기법

화이트-박스 기법은 공격자가 모델에 대한 모든 정보를 알고 공격하는 것

ex) FGSM(Fast Gradient Sign Method), PGD(Projected Gradient Descent)

원본 이미지

PGD기법
적용한 이미지
(eps = 0.1)

04

캡스톤 디자인
추진 전략

01. 예상 문제점
02. 진행사항
03. 역할 분담
04. 일정

04 추진전략

01. 예상 문제점

02. 진행사항

03. 역할 분담

04. 일정

예상 문제점

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

⇒ AI허브의 개방된 학습용 데이터셋을 사용

04 추진전략

진행사항

- 안티포렌식 기법이 적용된 데이터 셋 생성
- Xception 모델을 사용해서 real 이미지와 fake 이미지를 epoch 3으로 학습시킨 후, best model 저장

```
model = xception(num_out_classes=2, dropout=0.5)
print("=> creating model '{}'".format('xception'))
model = model.cuda(args.gpu)

fn = 'deepfake_c0_xception.pkl'
assert os.path.isfile(fn), 'wrong path'

model.load_state_dict(torch.load(fn))
print("=> model weight '{}' is loaded".format(fn))
```

```
----------
Epoch 1/3
Train:  0%|          | 0/227 [00:00<?, ?it/s]/usr/local/lib/python3.7/dist-packages/torch/utils/data/datalo
  cpuset_checked))
Train: 100%|██████████| 227/227 [08:50<00:00,  2.34s/it, loss - 0.0267, acc - 0.992]
Valid: 100%|██████████| 97/97 [01:31<00:00,  1.06it/s, loss - 0.0806, acc - 0.967]
Epoch 2/3
Train: 100%|██████████| 227/227 [07:52<00:00,  2.08s/it, loss - 0.0001, acc - 1.000]
Valid: 100%|██████████| 97/97 [01:13<00:00,  1.33it/s, loss - 0.0600, acc - 0.979]
Epoch 3/3
Train: 100%|██████████| 227/227 [07:54<00:00,  2.09s/it, loss - 0.0001, acc - 1.000]
Valid: 100%|██████████| 97/97 [01:13<00:00,  1.33it/s, loss - 0.0395, acc - 0.988]
```

04 추진전략

진행사항

- 저장한 모델을 불러와 inference 코드로 안티 포렌식 적용 이미지에 대한 결과 확인

```
model = xception(num_out_classes=2, dropout=0.5)
print("=> creating model '{}'".format('xception'))
model = model.cuda(args.gpu)

assert os.path.isfile(args.save_fn), 'wrong path'

model.load_state_dict(torch.load(args.save_fn)['state_dict'])
print("=> model weight '{}' is loaded".format(args.save_fn))

model = model.eval()
```

⇒ 원본 fake에 비해 안티 포렌식 적용 이미지에 대한 **성능하락을 확인**

```
1 print('-' * 50)
2 acc = validate(valid_loader, model, criterion)

--------------------------------------------------
Valid: 100%|██████████| 3100/3100 [17:15<00:00,  2.99it/s, loss - 0.8211, acc - 0.613]
```

04 추진전략

진행사항

- **앞으로의 계획 :**
  안티 포렌식 적용 이미지에 대한 탐지기 성능 하락을 확인하였으므로,
  원본 이미지와 안티 포렌식 적용 이미지를 사용하여 adversarial training 진행할 예정

04 추진전략

역할 분담

김지수
(팀장)

논문 및 자료조사
모델링

김민지

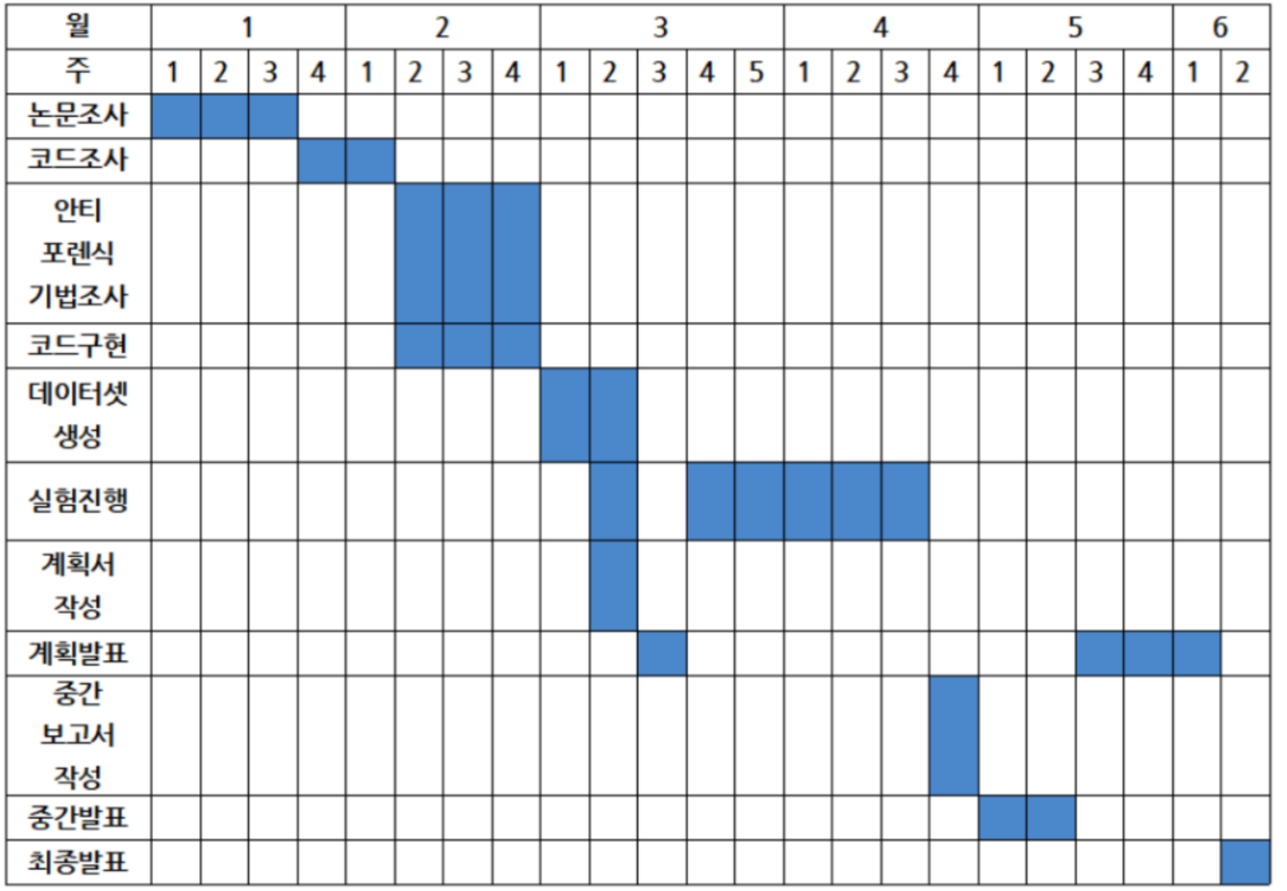
논문 및 자료조사
데이터 분석 및 전처리

민지민

논문 및 자료조사
모델 실험

04 추진전략

일정



감사합니다

20191769 김지수, 20191767 김민지, 20191730 민지민