다양한 유형의 안티포렌식에 강인한 딥페이크 탐지 모델 개발

팀명 : Attention

참여학생 : 김지수, 김민지, 민지민
지도교수 : 장한얼 교수님

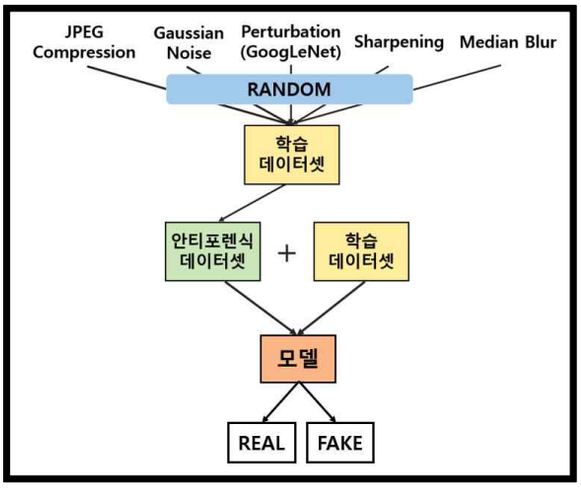
작품개요

딥페이크는 사회적으로 악영향을 끼치는 기술
딥페이크 사용의 96% → '포르노그라피'

기존 딥페이크 탐지기의 경우 적대적 공격뿐만 아니라 노이즈 추가, 선명화, JPEG 압축과 같은 간단한 이미지 변형 기법으로 쉽게 탐지가 우회된다는 문제점이 있다. 따라서 간단한 이미지 편집기법으로 우회되던 모델의 한계점을 적대적 학습을 통해 보완한 모델을 개발하였다.

작품 추진과정

• 기존 딥페이크 탐지 모델의 문제점인 이미지 편집 기법 기반 공격에 강인성을 획득하기 위한 방법론으로 적대적 학습 기법(Adversarial Training)을 선택했다. 적대적 학습 기법은 탐지를 우회하는 이미지를 학습에 포함하여 결국 공격 패턴을 학습하게 되어 강인성을 획득할 수 있는 방법이다.

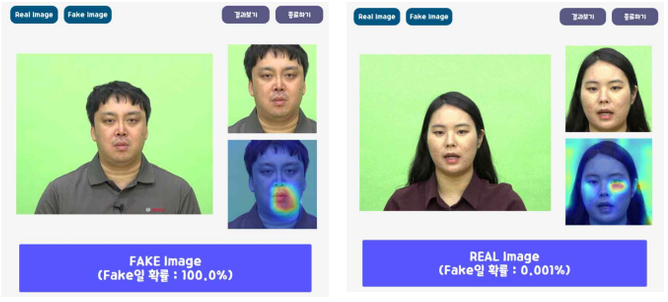


• 여러 가지 탐지 우회 공격 기법을 공격 순서와 공격 강도를 달리하여 랜덤하게 적용한 데이터를 생성해 적대적 학습에 반영하여 더 다양한 공격 패턴을 학습하게 되어, 학습에 반영하지 않은 공격에 대해서도 강인성을 획득하도록 했다.

최종 결과

	적대적 학습 전 정확도	적대적 학습 후 정확도
학습에 반영한 공격	80.6	92
학습에 반영하지 않은 공격	74.75	92.5

다양한 안티포렌식 기법이 random하게 적용된 데이터셋을 적대적 학습에 반영해서 학습한 딥페이크 탐지 모델로 학습에 반영하지 않은 안티포렌식 공격들에 대한 탐지율이다. 이를 통해 적대적 학습 기반의 딥페이크 탐지 모델이 학습에 반영하지 않은 안티포렌식 공격에도 우수한 성능을 보임을 확인할 수 있다.



탐지 모델을 적용한 GUI는 탐지할 때 필요한 얼굴 부분을 절삭한 부분과 탐지기가 어떤 부분을 보고 판단하였는지 알 수 있는 gradcam 이미지와 함께 하단에 탐지 결과와 fake 일 확률이 나오게 된다. gradcam 이미지는 색이 붉어질수록 결정에 큰 영향을 미치는 부분이다.

기대효과 및 활용방안

- 딥페이크 기술로 생성된 이미지 및 영상을 이용한 범죄 수사에 협조 가능
- 포토샵, 파이썬 등의 이미지 편집 도구로 안티포렌식을 수행한 딥페이크 이미지를 탐지
- 딥페이크 피해가 발생할 수 있는 곳이나 국가기관 및 군사기관 등 보안이 중요한 기업에 활용 가능