캡스톤디자인 I 계획서

<table>
<tr><td rowspan="2">제 목</td><td>국문</td><td colspan="4">안티포렌식에 강인한 딥페이크 탐지 모델 개발</td></tr>
<tr><td>영문</td><td colspan="4">Development of robust deepfake detector against anti-forensics</td></tr>
<tr><td>프로젝트 목표
(500자 내외)</td><td colspan="5">최근 딥페이크가 사회적으로 이슈가 되는 등 악영향을 끼치고 있는데, 이에 대해서 심각한 문제점을 느끼고 탐지에 대한 개선 필요성을 느꼈다. 사실 이에 대해 이미 탐지 기술은 많이 발전해있는 상태지만, 여러 논문을 보면 상당히 쉬운 우회기술을 통해서 딥페이크 탐지가 우회되고 있는 상황이다. 그래서 우리는 여러 탐지 우회기술인 안티포렌식 기법에도 강인한 딥페이크 탐지 모델 생성을 목표로 한다.</td></tr>
<tr><td>프로젝트 내용</td><td colspan="5">이 프로젝트의 궁극적인 목적은 안티 포렌식 데이터에도 강인한 모델을 만드는 것이다. 그러므로 먼저 딥페이크 데이터 셋을 구한 뒤 여러 블랙박스 공격(denoise, jpeg compression, sharpening, gaussian noise, salt and pepper noise)/화이트박스 공격(PGD/FGSM) 기법을 통해 안티 포렌식 데이터 셋을 직접 생성한다. 그 다음, 원본 데이터 셋과 생성한 안티 포렌식 데이터 셋을 함께 이용해서 adversarial training을 진행한 후 fine-tuning을 거쳐 우수한 성능의 모델을 생성한다.</td></tr>
<tr><td>중심어(국문)</td><td colspan="2">딥페이크</td><td colspan="2">안티포렌식</td><td>적대적 예제 생성 기법</td><td>적대적 훈련</td></tr>
<tr><td>Keywords (english)</td><td colspan="2">Deepfake</td><td colspan="2">Anti-forensics</td><td>Adversarial example generation method</td><td>adversarial training</td></tr>
<tr><td>멘토</td><td>소속</td><td colspan="2">해당사항 없음</td><td>이름</td><td colspan="2">해당사항 없음</td></tr>
<tr><td rowspan="4">팀
구성원</td><td>학년/반</td><td>·학 번</td><td colspan="2">이 름</td><td colspan="2">연락처(전화번호/이메일)</td></tr>
<tr><td>4</td><td>20191769</td><td colspan="2">김지수</td><td colspan="2">20191769@edu.hanbat.ac.kr</td></tr>
<tr><td>4</td><td>20191767</td><td colspan="2">김민지</td><td colspan="2">20191767@edu.hanbat.ac.kr</td></tr>
<tr><td>4</td><td>20191730</td><td colspan="2">민지민</td><td colspan="2">20191730@edu.hanbat.ac.kr</td></tr>
<tr><td colspan="7">컴퓨터공학과의 캡스톤디자인 관리규정과 모든 지시사항을 준수하면서 본 캡스톤디자인을 성실히 수행하고자 아래와 같이 계획서를 제출합니다.
2022 년 3월 11일
책 임 자 : 김지수 김지수
희망 지도교수 : 장한얼 교수님</td></tr>
</table>

1. 캡스톤디자인의 배경 및 필요성

최근 전 세계적으로 '가짜 뉴스' 및 '가짜 연예인 음란 동영상'에 사용되는 인공지능 기반의 딥페이크(Deepfakes) 기술이 사회적인 이슈로 대두되고 있다. 2019년 네덜란드의 보안업체 Deeptrace의 <The state of deepfakes 2019>에 따르면 전 세계 온라인 딥페이크 사용의 96%가 포르노그라피라고 한다. 심지어 해당 보고서에 따른 딥페이크 포르노그래피의 피해자 중 25%가 한국 여성이며 연예인과 일반인 모두 해당한다고 한다.

딥페이크 기술이란 딥러닝 기술을 이용해 악의적으로 조작된 음성, 영상, 이미지 등을 만들어 내는 방법이다. 인공지능 기술의 발전에 맞추어 더욱더 빠르고 정교한 생성 기술이 등장하고 있고, 이러한 딥페이크 기술은 빠른 개발 속도와 쉬운 접근성을 기반으로 '개인 정보 침해', '사기' 등 다양한 범죄에 악용되고 있다. 그에 따라, 딥페이크 영상 탐지에 대한 필요성이 대두되었고, 이런 영상 변조/합성 탐지에 대한 연구가 활발히 진행되고있다.

하지만, 이 연구의 가장 큰 문제점은 딥페이크 탐지기가 간단한 이미지 변형만으로도 탐지가 우회된다는 점이다. 그럼에도 불구하고, 딥페이크 우회 기술 탐지에 대한 연구는 현저히 적다. 이 점을 악용해 가해자들은 딥페이크 탐지를 우회하도록 이미지를 생성하고, 이를 인터넷 상에 퍼트려 범죄에 이용한다.

따라서 탐지가 어려운 딥페이크의 생성 및 확산을 방지하기 위한 우회 기법에 대한 고찰과 그렇게 생성된 이미지에 대한 탐지율 향상을 목표로 한 연구가 필요하다.

2. 캡스톤디자인 목표 및 비전

최근 딥페이크 탐지 기술이 이미 많이 발전해있는 상태지만, 여러 논문을 보면 이미지에 노이즈를 추가하거나 변형을 가하는 등 상당히 쉬운 우회기술을 통해서 딥페이크 탐지가 우회되고 있는 상황이다. 따라서 우리는 이러한 탐지 우회기술에 강인한 딥페이크 탐지 모델 생성을 목표로 한다.

이런 기술로 조작된 이미지나 영상을 좀 더 민감하게 판별함으로써 딥페이크 기술이 악용되지 않도록 할 수 있고, 더 나아가 관련된 범죄 발생률을 줄일 수 있다.

3. 캡스톤디자인 내용

이 프로젝트에서는 다양한 안티 포렌식 기법으로 생성된 데이터 셋에 대한 높은 딥페이크 탐지율을 도출할 수 있다. 이는 원본 데이터와 직접 생성한 안티 포렌식이 적용된 데이터 셋을 함께 이용해서 adversarial training 후 fine-tuning을 진행함으로써 얻을 수 있다.

직접 사용한 안티 포렌식 기법 즉, 적대적 예제 생성 기법에 대한 설명을 덧붙이자면, 크게 블랙박스 기법/화이트박스 기법이 있다. 블랙박스 기법은 신경망을 속이기 위해 이미지에 노이즈 추가하여 잘못 분류되도록 하는 방법이다. denoise, jpeg compression, sharpening, gaussian noise, salt and pepper noise 기법을 사용해 적대적 예제를 생성할 것이다.
화이트박스 기법은 모델의 gradient를 구한 후 손실을 증가시키는 방향으로 이미지가 잘못 분류될 때까지 노이즈를 추가하는 방법이다. FGSM(Fast Gradient Sign Method), PGD(Projected Gradient Descent) 공격 기법을 사용해서 적대적 예제를 생성할 계획이다.

상기 방식으로 이미지를 생성한 뒤, 관련 성능을 측정하는 실험을 진행한 다음 원본 데이터셋과 적대적 예제 생성 기법으로 생성한 데이터셋을 함께 학습해 적대적 예제에 대해서도 학습을 시켜 다양한 공격에도 강인한 모델을 만들 수 있다.

비 기능적 요구사항

보안 요구사항에 해당하는 부분으로 모델을 학습할 때 승인된 데이터 셋을 사용해야 한다. 그래서 AI 허브의 지능정보산업 인프라 조성 사업으로 추진되어 개방된 AI 학습용 데이터 셋을 사용할 계획이다.

4. 캡스톤디자인 추진전략 및 방법

컴퓨터 비전 프로젝트에서 자주 대두되는 문제는 데이터 셋의 저작권 문제이다. 그러므로 딥페이크 탐지라는 주제로 프로젝트를 진행한다면 가장 먼저 데이터 셋에 대한 문제를 마주할 것이라 예상했다. 더불어서 아무래도 사람의 얼굴에 관한 데이터를 사용해야 하다 보니 공개된 데이터 자체가 많지 않을뿐더러 초상권 문제도 발생할 수 있다. 이런 문제에 대비해 공인된 AI HUB의 데이터 셋을 주로 이용했다.

딥페이크 탐지의 소주제 선정을 위해 관련 논문 조사 및 실험을 진행하였고, 데이콘이나 캐글의 비슷한 주제의 개발물을 조사하고 직접 실행해 봤다.

학부 연구생 신분으로 ETRI(한국전자통신연구원)의 미세 신호 탐지 프로젝트에 참여한 경험이 있다. 그 프로젝트에서 이미지에 삽입된 미세 노이즈를 분석하는 역할을 맡아 육안으로 구별되지 않게 노이즈를 삽입하고, 그렇게 삽입된 노이즈를 분석하는 여러 방식에 대해 익혔다. 이 경험이 실험에 사용할 노이즈 데이터 셋을 생성하는 데 도움을 줬다.

프로젝트는 명확한 구분 없이 조원들끼리 매주 모여 함께 진행했다.

팀원명	역할
김지수	논문 및 자료조사, 모델 코드 작성
김민지	논문 및 자료조사, 모델 코드 작성
민지민	논문 및 자료조사, 모델 코드 작성 및 모델 실험