



## 캡스톤디자인 II 계획서

<b>제 목</b>	국문	적대적 학습을 통한 다양한 유형의 안티포렌식에 강인한 딥페이크 탐지 모델 개발		
	영문	Development of deepfake detection models that are robust to different types of anti-forensics through adversarial learning		
<b>프로젝트 목표 (500자 내외)</b>	<p>인공지능 기술이 발전하면서, 더욱더 빠르고 정교한 딥페이크 생성 기술이 등장하고 있다. 딥페이크 생성 기술은 ‘개인 정보 침해’, ‘사기’ 등 다양한 범죄에 악용되고 있고 딥페이크 생성방법이 고도화되어감에 따라 더욱 실제 같은 가짜 동영상, 이미지 등이 사회에 악영향을 미칠 수 있다. 이에 따라 다양한 딥페이크 탐지 기술이 연구 되었으며 현재 개발된 딥페이크 탐지 기술은 준수한 성능을 갖고 있다. 하지만 기존 탐지기의 경우 적대적 공격뿐만 아니라 additive noise, sharpening과 같은 간단한 이미지 변형에 의해서도 쉽게 탐지가 우회된다. 따라서 이러한 다양한 유형의 안티 포렌식에도 강인한 딥페이크 탐지 모델을 개발하는 것을 목표로 한다.</p>			
<b>프로젝트 내용</b>	<p>이 프로젝트의 궁극적인 목표는 적대적 공격과 간단한 이미지 변형만으로 딥페이크 탐지가 우회되는 문제를 해결하기 위해 다양한 유형의 안티포렌식에 강인한 딥페이크 탐지 모델을 개발하는 것이다. 이는 다양한 유형의 안티포렌식 기법(‘JPEG Compression’, ‘Gaussian Noise’, ‘Perturbation(GoogleNet)’, ‘Sharpening’, ‘Median Blur’)의 강도와 순서가 랜덤하게 적용된 데이터셋을 생성하고 안티포렌식 데이터셋을 원본 데이터셋과 함께 학습하는 적대적 학습 기법을 통하여 개발할 수 있다.</p>			
<b>기대효과 (500자 이내) (응용분야 및 활용범위)</b>	<p>개발한 모델을 사용하면 조작된 이미지나 영상을 좀 더 민감하게 판별함으로써 딥페이크 기술이 악용되지 않도록 할 수 있고, 더 나아가 관련된 범죄 발생률을 줄일 수 있다.</p>			
<b>중심어(국문)</b>	딥페이크	안티포렌식	적대적 예제 생성기법	적대적 학습
<b>Keywords (english)</b>	Deepfake	Anti-Forensics	Adversarial example generation method	adversarial training
<b>멘토</b>	소속	이름		
<b>팀 구성원</b>	학년 /반	학 번	이 름	연락처(전화번호/이메일)
	4	20191769	김지수	20191769@edu.hanbat.ac.kr
	4	20191767	김민지	20191767@edu.hanbat.ac.kr
	4	20191730	민지민	20191730@edu.hanbat.ac.kr
<p>컴퓨터공학과와 캡스톤디자인 관리규정과 모든 지시사항을 준수하면서 본 캡스톤디자인을 성실히 수행하고자 아래와 같이 계획서를 제출합니다.</p> <p style="text-align: center;">2022 년    7    월    1    일</p> <p style="text-align: right;">책 임 자 : 김지수       지도교수 : 장한얼      </p>				

# 캡스톤디자인 계획서(양식)

## 1. 캡스톤디자인의 배경 및 필요성

- 1) 수행하려는 프로젝트 과제와 관련되는 국내·외 연구, 산업 현황, 문제점 및 전망 등에 관하여 기술
- 2) 프로젝트의 필요성을 구체적으로 기술

딥페이크 기술은 딥러닝을 통해 사람의 얼굴 특징을 분석한 후 매우 짧은 시간에 이미지를 합성할 수 있는 기술이다. 딥페이크 기술을 활용하여 돌아가신 가족을 실제 살아있는 것처럼 재현할 수 있고, 여러 사람의 얼굴을 통해 가상의 얼굴을 제작해서 모자이크 대신 얼굴을 합성할 수도 있다. 하지만, 최근 전 세계적으로 동영상 합성 동영상에 사용되는 인공지능 기반의 딥페이크(deepfakes) 기술은 ‘개인 정보 침해’, ‘사기’, ‘음란물 생성’ 등 다양한 범죄에 악용되고 있다.

2019년 네덜란드의 보안업체 Deeptrace의 ‘The state of deepfakes 2019’에 따르면 전 세계 온라인 딥페이크 사용의 96%가 포르노그래피라고 한다. 심지어 딥페이크 포르노그래피의 피해자 중 25%가 한국 여성이며 연예인뿐만 아니라 일반인 모두 해당된다. SNS에 일상을 올렸을 뿐인데 피해를 받고 있으며 공격자는 딥페이크로 만든 포르노 영상을 유포하겠다고 협박뿐만 아니라 금품까지 요구한다. 국내에서도 N번방 사건에서 딥페이크를 악용한 사진이 거래 및 유포된 정황이 확인되었다. 심지어 최근 영국에서는 딥페이크로 만들어진 음성을 이용한 보이스피싱 범죄까지 나오고 있다. 이처럼 딥페이크가 악의적인 목적으로 활용될 수 있기 때문에 딥페이크 영상인지 아닌지 판별하는 기술은 매우 중요하다. 딥페이크 기술이 빠른 개발 속도와 쉬운 접근성을 기반으로 더 고도화됨에 따라 더욱 심각한 사회적 문제를 야기할 것이다. 이에 따라 다양한 딥페이크 탐지 기술이 연구되었으며 현재 개발된 딥페이크 탐지 기술은 준수한 성능을 갖고 있다.

하지만 기존 딥페이크 탐지기는 적대적 공격뿐만 아니라 노이즈 추가, 선명화, JPEG 압축과 같은 간단한 이미지 변형 기법으로 쉽게 탐지가 우회된다는 문제점이 있다. 이러한 딥페이크 탐지 기법과 같은 디지털 포렌식을 회피하는 기술을 안티포렌식이라고 한다. 따라서 딥페이크 기술이 점점 더 정교해지고 안티포렌식 사례가 등장함에 따라 다양한 안티포렌식 유형에 강인한 딥페이크 탐지 연구가 필요한 상황이다.

## 2. 캡스톤디자인 목표 및 비전

- 1) 현안 해결, 기존지식 개선, 기존 원리의 새로운 규명, 새로운 원리에 기반한 차세대 지식, 완전히 새로운 발견/발명 등을 중심으로 수행 프로젝트의 창의성·도전성을 기술함

인공지능 기술이 발전하면서, 더욱더 빠르고 정교한 딥페이크 기술이 등장하고 있다. 딥페이크 생성 방법이 고도화되어감에 따라 이에 대응하는 다양한 딥페이크 탐지 기술이 연구되었으며 현재 개발된 딥페이크 탐지 기술은 준수한 성능을 갖고 있다. 하지만 기존 딥페이크 탐지 기술은 universal perturbation과 같은 적대적 공격 뿐만 아니라 noise, sharpening, JPEG compression등과 같은 간단한 이미지 편집만으로도 쉽게 우회가 되는 문제가 발생한다. 따라서 이러한 다양한 유형의 안티포렌식에 강인한 딥페이크 탐지모델을 개발하는 것을 목표로 한다.

1) 캡스톤디자인의 주요 기능, 비 기능적 요구사항(성능, 보안, 유지보수성 등)

이 프로젝트에서는 다양한 유형의 안티포렌식 기법으로 생성된 데이터셋에 대한 높은 디펙이크 탐지율을 도출할 수 있다. 이는 다양한 유형의 안티포렌식 기법이 적용된 데이터셋을 생성하고 안티포렌식 데이터셋과 원본 데이터셋을 함께 학습하는 적대적 학습 기법을 통하여 얻을 수 있다.

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

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

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																					

#### 4. 캡스톤디자인 추진전략 및 방법

##### 1) 캡스톤디자인 목표 달성을 위한 추진전략, 수행방법 및 추진절차를 기술함

##### **\* 참고: 추진전략의 작성 \***

- (1) 캡스톤디자인에 대한 이해  
캡스톤디자인 추진을 위한 예상 문제점을 식별하고 프로젝트 추진을 위한 준비방안 수립
- (2) 캡스톤디자인 경험  
저학년 프로젝트를 통해 습득한 기본 기술 활용 및 Lessons Learned를 활용
- (3) 검증된 멘토 활용  
관련된 경험과 역량을 풍부하게 축적하고 있는 멘토를 활용
- (4) 프로젝트 관리체계 수립  
역할 및 책임을 명확히 하도록 프로젝트 관리 체계를 수립

컴퓨터비전 프로젝트에서 자주 대두되는 문제는 데이터셋의 저작권 문제이다. 그러므로 딥페이크 탐지라는 주제로 프로젝트를 진행한다면 가장 먼저 데이터셋에 대한 문제를 마주할 것이라 예상했다. 더불어서 아무래도 사람의 얼굴에 관한 데이터를 사용해야 하다보니 공개된 데이터 자체가 많지 않을 뿐더러 초상권 문제도 발생할 수 있다. 이런 문제에 대비해 공인된 AI HUB의 데이터셋을 주로 이용했다.

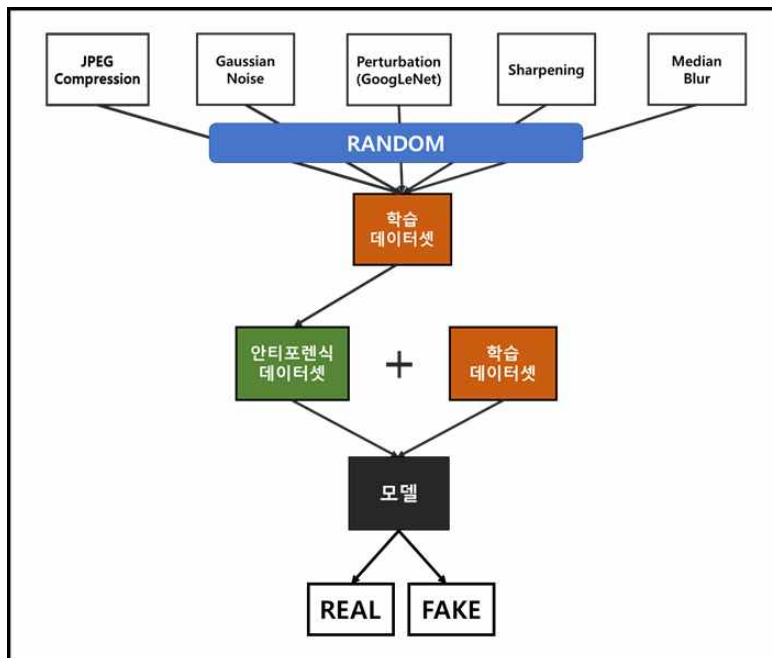
학부연구생 신분으로 정부출연연구소의 미세신호 탐지 프로젝트에 참여한 경험이 있다. 그 프로젝트에서 이미지에 삽입된 미세 노이즈를 분석하는 역할을 맡아 육안으로 구별되지 않게 노이즈를 삽입하고, 그렇게 삽입된 노이즈를 분석하는 여러 방식에 대해 익혔다. 이 경험이 실험에 사용할 노이즈 데이터셋을 생성하는데 도움을 줬다.

프로젝트는 명확한 구분 없이 조원들끼리 매주 모여 함께 진행했으며, 각자 주어진 환경에 맞추어 업무를 부담했다.

팀 구성	김 지 수	김 민 지	민 지 민
역할	실험 데이터셋 생성	학습 모델 생성	학습 모델로 성능 확인

**\* 참고: 수행방법의 작성 \***

캡스톤디자인 수행을 위한 방법론, 프레임워크, 분석틀에 대해 작성



**<수행방법>**

위의 그림과 같은 모델 학습 기법을 통하여 다양한 종류의 안티포렌식에 강인한 모델 개발을 목표로 한다. 먼저, 원본 데이터셋만을 사용해 딥페이크 탐지기를 학습한 Xception기반 모델을 도출한다. 그 다음, 안티포렌식 데이터셋을 생성한다. 안티포렌식 기법으로 적용할 이미지 편집 기법으로는 median blur, gaussian noise, sharpening, JPEG compression, universal perturbation 으로 이 공격들의 강도와 순서가 랜덤으로 적용된 데이터셋을 생성한다. 마지막으로, 원본 데이터셋과 강도 및 공격 순서가 랜덤하게 적용된 안티포렌식 데이터셋을 학습 데이터셋으로 사용하는 적대적 학습을 통해 최종 모델을 도출해낼 것이다.

**2) 캡스톤디자인 목표 달성을 위한 팀 구성 체계 및 역할에 대하여 기술함**

팀 구성	김 지 수	김 민 지	민 지 민
역할	실험 데이터셋 생성	학습 모델 생성	학습 모델로 성능 확인

**5. 캡스톤디자인 결과의 활용방안**

**1) 제안된 캡스톤디자인이 추진되었을 경우의 사회적/기술적/경제적 파급효과 등을 자유롭게 기술함**

딥페이크 기술은 ‘개인 정보 침해’, ‘사기’ 등 다양한 범죄에 이용되고 있다. 저희가 만든 모델을 이용하면, 직접적으로 범죄가 발생하지 않도록 할 수는 없지만 저희가 만든 모델을 이용하면 피해자가 피해 본 사실을 밝힐 수 있다. 즉, 딥페이크로 합성된 이미지인지 아닌지 판별해냄으로써 관련 범죄

발생률을 조금이나마 줄일 수 있다.

현재 사용되고 있는 딥페이크 탐지기가 초보자도 쉽게 할 수 있는 이미지 편집기법 공격으로 무력화되고 있는데, 저희가 만든 모델은 이 점에 초점을 두어 안티포렌식으로 조작된 이미지나 영상을 좀 더 민감하게 판별함으로써 딥페이크 기술이 악용되지 않도록 도울 수 있다.

## 6. 참고문헌

## 캡스톤디자인 II 계획발표 채점표

팀 구성원	학년/반	학 번	이 름				
제 목							
항목			점수				
			1	2	3	4	5
1. 프로젝트 주제의 필요성이나 중요성이 적절히 서술되었는가?							
2. 국내외 동향(문제 제기), 주요 기능(특징 포함) 및 범위가 적절히 서술되었는가?							
3. 기대효과(사회적, 기술적, 경제적 파급효과)가 적절히 서술되었는가?							
4. 추진 전략과 수행방법이 적절한가?							
5. 팀 구성과 역할 분담이 적절히 이루어졌는가?							
합계							
*수정 및 개선 의견							
<div style="text-align: center;">2013년    월    일</div> <div style="display: flex; justify-content: space-between; margin-top: 20px;"> <span>심사위원 :</span> <span>(인)</span> </div>							

※ 채점은 각 영역별 5점 만점을 기준으로 채점함.(상 5, 중 3, 하 1)

※ 계획서와 발표내용을 참고하여 채점표에 따라 평가함.