DU-도전학기 결과보고서

과제명	Anti-성착취 딥페이크 기술 개발		
참여자	성명	소속	학번
		컴퓨터공학전공	
		컴퓨터공학전공	
		컴퓨터공학전공	
		컴퓨터공학전공	
지도교수 의견	상기 학생들은 기존 딥페이크 탐지 연구에서 다루지 않았던 성범죄 분야에 특화된 딥페이크 탐지 연구를 한 학기 동안 주도적으로 수행하였습니다. 공개된 성범죄 딥페이크 데이터셋이 존재하지 않은 관계로, 학생들은 직접 크롤러를 개발하여 3만 개 이상의 데이터를 수집하였고, CNN 기반 모델과 그래프 신경망(GAT)을 결합한 하이브리드 딥러닝 모델을 통해 고도화된 탐지 기술을 구현하였습니다. 해당 연구 결과는 2025년 한국멀티미디어학회 춘계 학술대회에서 발표되었으며 기술적 완성도와 우수성을 인정받아 좌장 추천으로 KCI 등재지 확장 논문으로 추천되었습니다. 그러나, 학생들이 개발한 모델을 조금 더 확장하면 SCI급 논문으로도 게재할 수 있을 것이라 판단하여 여름방학동안 심화 연구를 진행할 계획입니다. 이러한 성과는 학생들이 주 2~3회 이상 오프라인 회의를 통해 지속적으로 소통하고 협력한 결과이며 목표 달성을 위해 끊임없이 고민하고 열정을 다한 노력의 결실이라고 생각합니다. 또한, 학생들이 주도적으로 참여할 수 있는 도전학기 프로그램을 운영해 주신데에 깊이 감사드립니다. (소속) 컴퓨터정보공학부		

1. 도전 과제 내용

딥페이크는 딥러닝과 가짜의 합성어로 GAN 기술이 사용된 인공지능을 기반으로 인간 합성 이미지 및 영상을 제작하는 기술이다. 이 딥페이크 기술은 최근 들어 표정, 각도, 조명, 피부 질감까지 자연스럽게 합성이 가능할 정도로 고도화되어 육안으로 식별이 어려운 수준에 도달했다. 영화 합성, 인공지능 교육, 영상 편집 등의 기존의 긍정적인 목적과 달리 성범죄 목적의 악용 사례가 급증하며 사회적인 문제로 대두되고 있다. 하지만 기존에 이루어진 딥페이크 탐지 모델 개발 연구들은 성범죄 딥페이크와 거리가 먼 공개 딥페이크 데이터셋을 사용하였으며, 직접 학습하여 모델을 생성한 결과, 실제로 웹사이트 및 소셜미디어에서 유포되는 성범죄 딥페이크물을 탐지하지 못하는 것을 확인하였다. 본 팀은 이러한 문제를 해결하기 위해 웹사이트와 소셜미디어에서 직접 크롤링하여 데이터셋을 수집하고 이를 기반으로 CNN과 GAT를 결합한 하이브리드 모델을 개발하여 성범죄 딥페이크 탐지에 최적화된 ‘하이브리드 모델 기반 성범죄 딥페이크 탐지 시스템’을 제안하기 위한 도전학기 프로젝트를 진행하였다.

본 팀은 먼저 데이터셋을 수집하기 위하여 딥페이크 성범죄 이미지와 일반 성범죄 이미지를 유포하는 웹페이지 및 소셜미디어(Telegram, X(구 Twitter), Pinterest)에서 데이터를 수집하였다. 수집한 이미지는 웹페이지에서 렌더링하여 딥페이크의 종류(얼굴 합성, 신체 변형, 기타 합성 등)을 수동으로 분류하였으며, 이를 기반으로 이미지를 주파수 영역으로 변환하여 저주

파・고주파의 특징점을 강조해주는 고속 퓨리에 변환(Fast Fourier Tranform, FFT)과 이미지 영역 기반으로 얼굴・신체・배경 등의 분리를 통해 각각의 시각적 패턴을 벡터값으로 변환하여 시각적 특징 벡터를 추출하는 데이터 전처리를 수행하여 학습에 효율성을 높였다. 이후 CNN을 기반으로 이미지 내 합성곱 계층 및 풀링 계층을 통해 시각적 특징 벡터를 추출하고, 추출된 벡터를 그래프 신경망 기반의 GAT를 통해 영역 간의 상호 구조적 관계성을 학습하여 CNN-GAT 하이브리드 기반 딥페이크 탐지 모델을 생성하였으며 웹페이지 및 소셜미디어에서 성범죄 딥페이크 탐지가 가능한 시스템을 구현하였다.

2. 도전 과제 수행 결과 및 성과

본 팀은 먼저 데이터셋 수집을 위해 웹페이지 및 소셜미디어에서 직접 크롤링을 통하여 데이터셋을 수집하였다. 일반 성범죄물 유포 사이트 3곳과 딥페이크 성범죄물 유포 사이트 3곳에서 딥페이크 성범죄 이미지와 일반 성범죄 이미지 각 10,000장씩 수집하였으며, 소셜미디어에서는 텔레그램, X, Pinterest에서 딥페이크 성범죄 이미지와 일반 성범죄 이미지를 각 5,000장을 수집하였다. 이 데이터셋을 기반으로 CNN에서 이미지 내 얼굴, 신체, 배경 등을 인식하여 각 부분의 시각적 특징 벡터값을 추출하여 노드로 지정하고, 각 노드 간의 벡터를 그래프 신경망 기반의 GAT를 통해 노드 간의 상호 구조적인 관계를 학습하여 성범죄 딥페이크를 탐지할 수 있는 모델을 개발하였다. 이 하이브리드 모델을 비교하기 위한 단일 CNN 및 단일 GAT 모델을 개발한 후 비교한 결과, 하이브리드 모델이 가장 정확한 인식률을 보여주는 것을 확인하였다. 이후 웹 서버를 개발하여 실제로 누구나 딥페이크 이미지를 탐지할 수 있으며 추가적으로 신고 웹페이지 연결 및 세부적인 기능들을 추가하였다.

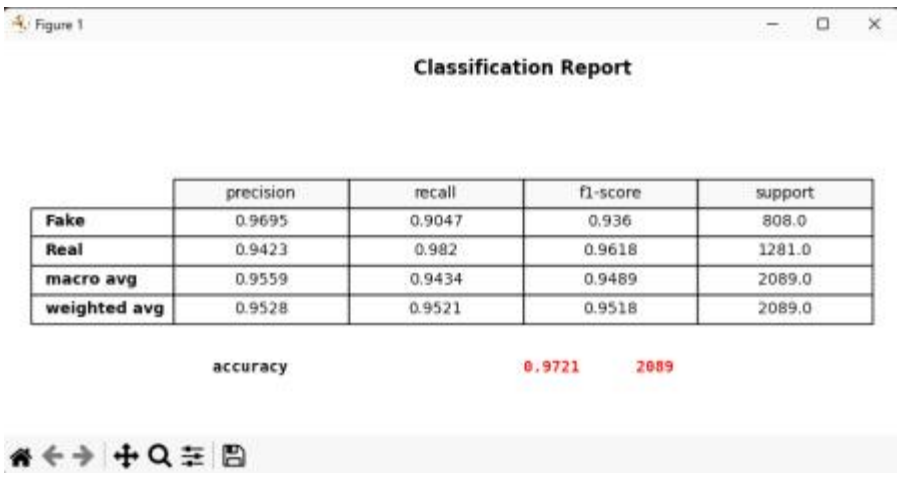

그림 1. GAT-CNN 하이브리드 모델 딥페이크 탐지 정확도

표 1. CNN 및 GAT 단일·결합 모델의
딥페이크 성범죄물 탐지 정확도 결과

구분	모델	Accuracy
단일	CNN	95.5%
	GAT	80.3%
하이브리드	CNN-GAT	97.2%

3. 자기 평가

- 팀원 개별 과제 수행 결과

- **강00** : 팀장 역할을 맡았으며 웹사이트 및 소셜미디어에서 딥페이크/일반 성범죄 이미지 데이터를 크롤링하였으며 이후 퓨리에 전처리를 진행하여 데이터 수집 및 전처리를 진행하였다. 또한 팀원들과 함께 한국멀티미디어 학회에 논문을 투고하기 위하여 논문의 전체적인 부분을 작성하였으며 춘계학술대회에서 발표를 담당하였다. 이후 웹페이지 제작 및 딥페이크 탐지 모델 학습 등 다른 팀원의 역할에 모두 참여하며 프로젝트가 잘 수행되도록 함으로써 성공적으로 마무리하였다.

- **김00** : 해당 프로젝트에서 성범죄물에 특화된 GNN 기반 딥페이크 탐지 모델 개발을 맡았다. 처음 계획과는 달리 전처리 여부에 따른 모델 개발이 아닌, 단일 GNN 모델과 CNN-GNN 하이브리드 모델 개발로 방향을 수정하였다. 이때, 가중치를 기반으로 중요 노드와 간선을 학습할 수 있는 GAT 모델을 적용하여 각각 단일 GAT 모델과 CNN-GAT 하이브리드 모델을 구현하였으며, 이후 모델 성능 평가를 진행하였고, 한국멀티미디어학회에 투고한 학술지의 일부 작성 및 발표 자료 제작 등 팀장 보조 역할도 성실히 수행하였다.

- **김00** : 프로젝트에 대한 선행 연구를 찾아보기 위하여 관련 논문 및 연구 등을 사전 조사하여 기존에 존재하는 딥페이크 탐지 모델 및 학습을 위한 데이터셋, 전처리 기법 및 인식률 증가를 위한 방법 등을 찾아보고 어떤 방향으로 진행할지에 대한 정보를 정리하였으며, 단일 CNN 모델을 제작하며 소셜미디어에서 유포되는 딥페이크 성범죄 이미지 링크를 수집하기 위한 키워드를 선정하였으며, 논문 투고를 위해 관련 연구 부분을 작성하였다.

- **박00** : 웹사이트에서 음란물 정상 이미지와 정상 데이터 셋을 크롤링 하였으며, 학습에 방해가 생기는 데이터를 따로 삭제하여 딥페이크 탐지 모델의 학습 정상 이미지 데이터셋을 구축하였다. 또한 한국 멀티미디어 학회 논문 투고를 위해 데이터셋 구축과 데이터셋 전처리 부분을 작성하였다. 이후 개발된 딥페이크 모델을 활용해 웹페이지를 제작하였으며, 이때 라즈베리파이를 이용하여 서버를 구동시켜 웹 사이트 개발을 하였다.

- 초기에 계획한 연구 중 여러 가지 문제를 직면하여 다른 방법을 찾아보고 계획을 변경하거나 데이터 전처리, 모델 학습, 탐지 시스템 개발까지에 있어서 많은 어려움이 존재했지만, 팀원 모두 끝까지 포기하지 않고 팀워크를 맞춰 잘 진행하였기에 **‘성범죄 딥페이크 탐지 기술 개발’** 프로젝트 진행을 성공적으로 완료하였다.

4. 최종 결과물

- 개인(팀원별) 결과물

- **강00** : Requests 라이브러리, Telegram API를 이용하여 웹페이지 6곳 및 각종 소셜미디어에서 성범죄/일반 딥페이크 데이터셋 크롤링 및 고속 퓨리에 변환 데이터 전처리 진행 및 웹서버 제작 및 논문 작성

그림 2. 일반/딥페이크 성범죄 이미지 데이터 수집 워크플로우

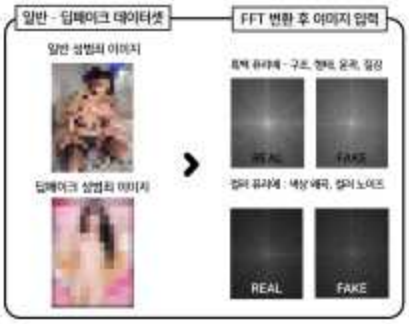

그림 3. 이미지 내
고주파/저주파를 강조하는 퓨리에
변환 전처리 결과

- **김00** : GAT 단일 모델 구축 및 성범죄 이미지 내 얼굴, 신체, 배경 영역을 자동으로 인식하여 벡터값을 지정해주는 시각적 특징 벡터 추출 및 최종적인 CNN-GAT 하이브리드 모델 개발 및 논문 작성

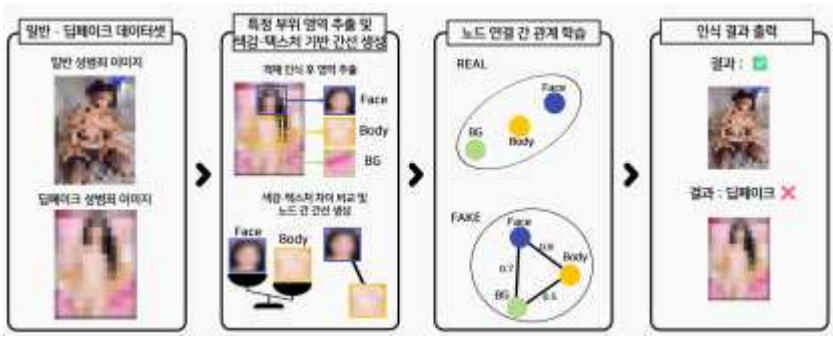

그림 4. GAT 단일 모델 학습 및 모델 구축 워크플로우

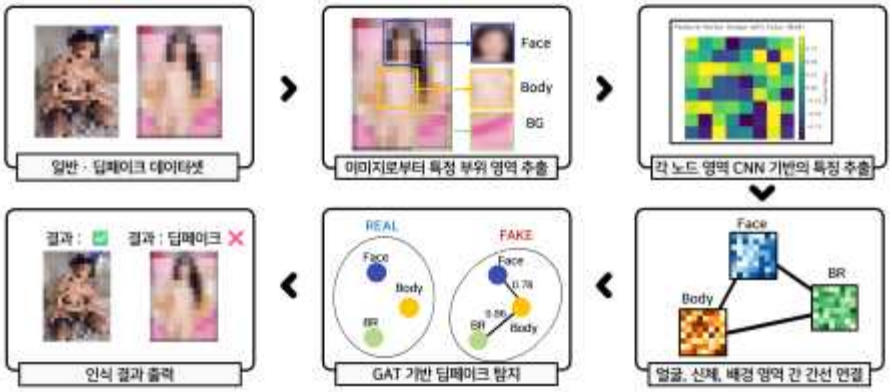

그림 5. GAT-CNN 하이브리드 모델 학습 및 모델 구축 워크플로우

- 김00 : CNN 단일 모델 구축 및 이미지 특징을 자동으로 강조해주는 CBAM attention 알고리즘 구현, 기존 딥페이크 탐지 연구 동향(모델 생성, 인식 효율 증가 방법, 학습 데이터셋 등) 조사 및 논문 작성

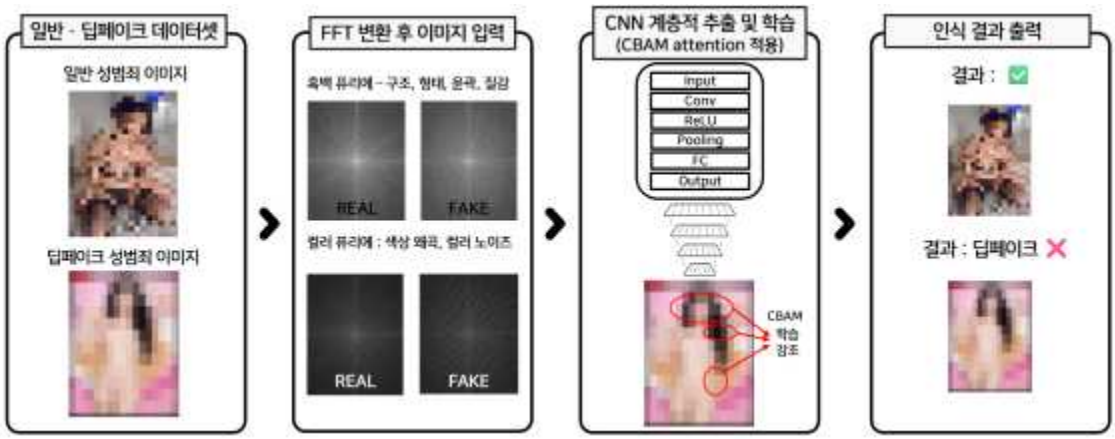

그림 7. 단일 CNN 모델 학습 및 모델 구축 워크플로우 및 CBAM attention 적용 알고리즘 구현

Reference	Dataset	Criminal Type	
		General	Porno
[4]	- FaceForensics++ - DFDC - Celeb-DF - UADFV - Google Deepfake Detection - DeepFake-TIMIT - DeeperForensics-1.0 - DF-VIVID - DFDC-preview - WildDeepfake	○	
[5]	- FaceForensics++ - Celeb-DF (v2)	○	
[6]	- CelebA + FFHQ - AttGAN - GDWCT - StyleGAN, StyleGAN2 - StarGAN	○	
[7]	- FaceForensics++ - Celeb-DF (v2)	○	
[8]	- FaceForensics++ - DeepFaceLab - FaceShifter	○	
Our Model	- a self-constructed dataset	○	○

그림 6. 기존 딥페이크 탐지 연구 동향 조사

- **박00** : Requests 라이브러리를 이용하여 일반/딥페이크 성범죄 이미지 크롤링 및 딥페이크 종류 수동 분류, 이미지 학습을 위한 기본적인 전처리 진행 및 시각적 특징 벡터 추출 전처리 진행, 웹페이지 제작 및 라즈베리파이를 이용하여 웹 서버 구동 및 논문 작성

그림 8. 각종 학습에 효율 증가를 위한 이미지 전처리 기법 수행

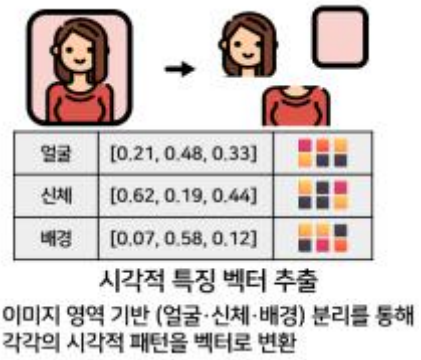

그림 9. 이미지 영역별 시각적 특징 벡터 추출 전처리

- 팀 공통 결과물

그림 10 딥페이크 탐지 시스템 서버 및 웹페이지 구현

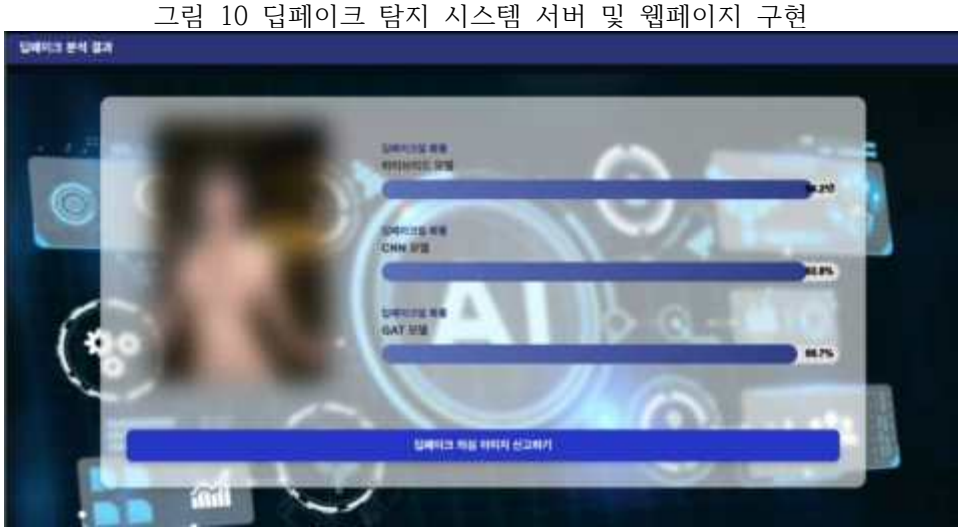

그림 11. 딥페이크 성범죄 이미지 탐지 결과 렌더링 및 신고 버튼 기능

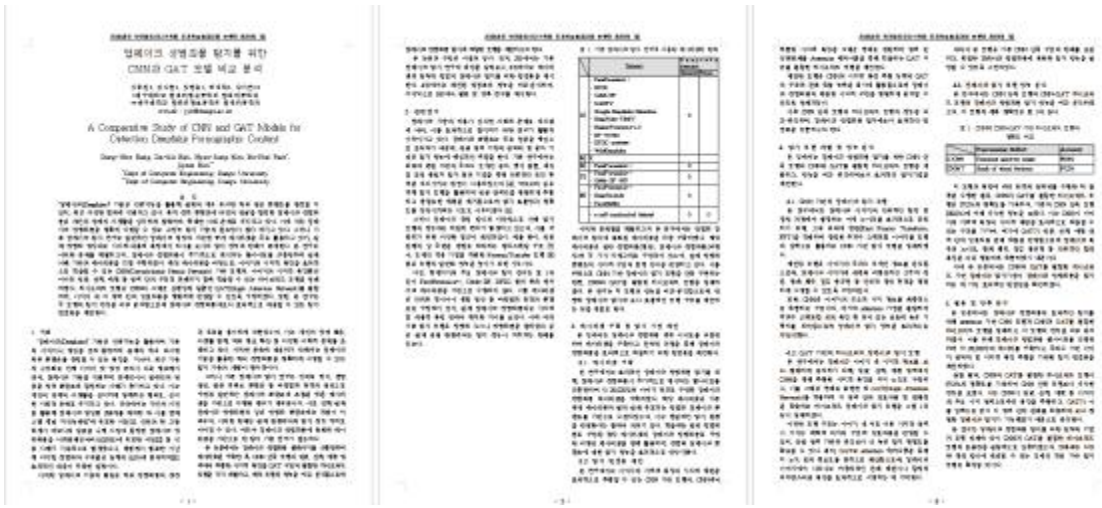

그림 12 한국멀티미디어 학회 투고 학술지

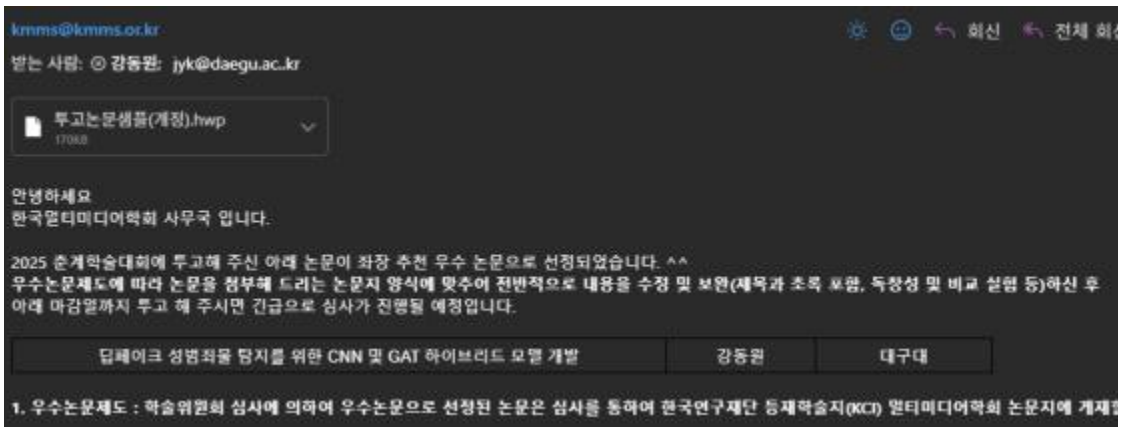

그림 14. 한국멀티미디어학회 좌장 추천 우수 논문 선정