딥페이크 사전 보호 및 탐지를 위한 모바일 시스템

김다영[1], 조영준[2]
[1]전남대학교 인공지능융합학과 석사과정
[2]전남대학교 인공지능융합학과 부교수
qetuo090909n@jnu.ac.kr, yj.cho@jnu.ac.kr

Mobile System for Deepfake Prevention and Detection

Da-Yeong Kim[1], Yeong-Jun Cho[2]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

최근 딥러닝 기반 생성 모델의 발전으로 인해 딥페이크 영상과 이미지가 급속히 확산되고 있으며, 이는 개인의 프라이버시와 사회적 신뢰에 심각한 위협을 초래한다. 본 논문에서는 사용자가 일상 속에서 손쉽게 활용할 수 있는 딥페이크 사전 보호 및 탐지 모바일 시스템을 제안한다. 제안 시스템은 이미지 및 동영상 입력에 대해 얼굴 영역을 검출하고, 합성 여부를 판별하는 딥러닝 기반 탐지 모델을 제공한다. 또한 사후 대응 차원에서 위조 여부를 탐지하는 것에 그치지 않고, 콘텐츠 업로드 전 사전 보호용 노이즈 삽입 기능을 제공하여 딥페이크 생성에 악용되는 것을 방지한다. 본 논문에서는 시스템의 전반적인 구성과 핵심 알고리즘, 다중 이미지 및 영상 처리 방식, 그리고 사전 보호 기능의 효과를 상세히 설명한다.

1. 서론

딥러닝 기반 생성 모델(GAN, Diffusion 모델 등)의 발전으로 인해 사람의 얼굴과 음성을 사실적으로 합성하는 것이 가능해졌다. 이러한 기술은 영상 제작이나 예술적 창작 분야에서 긍정적으로 활용되기도 하지만, 동시에 범죄나 허위 정보 유포와 같은 부정적 목적으로 악용되는 사례가 급격히 늘고 있다. 특히 딥페이크는 SNS, 메신저, 영상 플랫폼 등을 통해 빠르게 확산되며 개인의 명예 훼손, 성범죄, 정치적 조작 등의 사회적 문제를 유발한다.

기존 연구들은 대부분 생성된 콘텐츠를 판별하는 사후 탐지(Post-detection) 방식에 집중하였다. 그러나 실제 사회적 피해를 최소화하기 위해서는 사후 탐지뿐만 아니라, 애초에 위조 콘텐츠가 생성되지 않도록 막는 사전 보호(Prevention) 기법이 필요하다. 사용자가 이미지를 업로드할 때 미세한 노이즈를 삽입하여 학습에 적합하지 않은 데이터를 제공함으로써, 딥페이크 합성 모델이 활용할 수 없도록 만드는 방식이 대표적인 예이다.

이에 본 논문에서는 사용자가 모바일 환경에서 쉽게 활용할 수 있는 시스템을 목표로 한다. 제안 시스템은 두 가지 주요 기능을 제공한다. 첫째, 이미지 및 동영상을 입력받아 딥페이크 여부를 판별하는 탐지 기능을 제공한다. 둘째, 사용자가 SNS나 메신저에 이미지를 업로드하기 전에 사전 보호용 노이즈를 자동 삽입하여, 해당 이미지가 딥페이크 학습에 악용되지 않도록 한다.

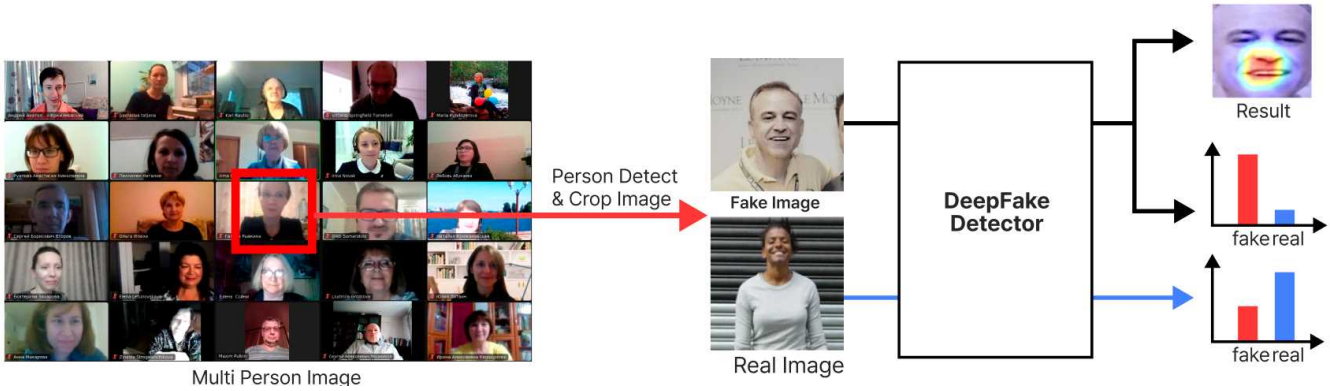
2. 제안방법

2.1 시스템 개요

제안 시스템은 모바일 애플리케이션 형태로 구현되며, 프론트엔드(React Native 기반)와 백엔드(Django 기반), 그리고 딥러닝 탐지 모델(Pytorch 기반)로 구성된다. 사용자가 앱에 이미지나 영상을 업로드하면, 서버는 이를 프레임 단위로 처리한 후 얼굴 영역을 검출하고 딥페이크 여부를 판별한다. 또한 사용자는 업로드 전 사전 보호 기능을 활성화하여 노이즈가 삽입된 안전한 콘텐츠를 공유할 수 있다.

2.1 다중 이미지 딥페이크 탐지

이미지 입력의 경우, 시스템은 다음과 같은 절차를 수행한다. 얼굴 검출 및 정규화: 입력된 이미지에서



(그림 1) 딥페이크 탐지 시스템

얼굴 영역을 탐지한 뒤 일정한 크기로 정규화한다.

특징 추출: 합성 과정에서 발생하는 미세한 아티팩트를 포착하기 위해 CNN 기반 특징 추출기[1]를 적용한다.

분류 단계: 추출된 특징 벡터를 분류기에 입력하여 진짜(real) 또는 가짜(fake) 여부를 판별한다.

결과 제공: 최종 결과는 확률 기반 신뢰도 점수와 함께 사용자에게 제공된다.

모델 학습 과정에서는 실제(real)와 합성(fake) 이미지의 차이를 반영하기 위해 크로스 엔트로피 손실 함수를 사용한다. 손실 함수는 다음과 같이 정의된다.

$$\mathcal{L} = -\sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right]$$

위 수식은 모델이 예측한 합성 확률이다.

2.3 동영상 딥페이크 탐지

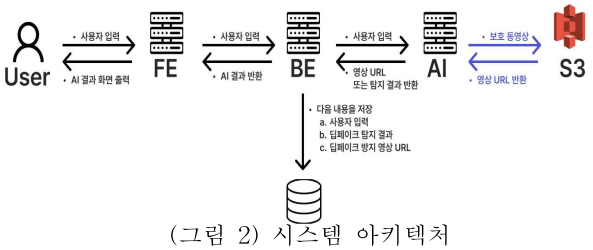
동영상 입력의 경우, 영상 전체를 처리하는 대신 효율성을 위해 일정 간격(초당 3~5프레임)으로 프레임을 추출한다. 각 프레임은 이미지 탐지와 동일한 절차를 거쳐 진위 여부가 판별된다. 이후 프레임 단위 결과를 집계(majority voting)하여 최종적으로 해당 영상이 합성된 것인지 여부를 결정한다.

이 방식은 일부 구간만 합성된 경우에도 탐지가 가능하며, 프레임 누락이나 화질 저하가 있는 환경에서도 안정적으로 동작한다. 결과적으로 영상 전체에 대한 딥페이크 여부를 사용자에게 직관적으로 제공한다.

2.4 사전 보호용 노이즈 삽입

본 시스템의 가장 큰 특징은 사후 대응이 아닌 사전 보호 기능을 제공한다는 점이다. 사용자가 SNS나 메신저에 이미지를 업로드할 때, 앱은 얼굴 영역에 미세한 adversarial noise[2]를 삽입한다. 이 노이즈는 인간의 눈으로는 거의 구별할 수 없지만, 합성 모델이 학습할 경우 성능 저하를 유발한다. 이를 통해 사용자의 이미지가 딥페이크 합성 데이터로 악용되는 것을 사전에 차단한다.

사전 보호 기능은 사용자가 선택적으로 활성화할 수 있으며, 개인 프라이버시와 콘텐츠 안전성을 동시에 강화한다.



(그림 2) 시스템 아키텍처

4. 결론

본 논문에서는 딥페이크 사전 보호 및 탐지를 위한 모바일 시스템을 제안하였다. 제안 시스템은 이미지와 영상 입력 모두에 대해 얼굴을 검출하고 합성 여부를 판별하며, 실시간으로 결과를 제공한다. 또한 기존 연구들과 달리 사후 탐지에만 머무르지 않고, 사전 보호용 노이즈 삽입 기능을 통해 사용자가 업로드하는 콘텐츠가 악용되는 것을 원천적으로 방지한다.

본 시스템은 누구나 모바일 환경에서 손쉽게 사용할 수 있으며, 개인의 디지털 권리를 보호하고 딥페이크 확산을 억제하는 데 기여한다. 향후에는 다양한 공격 환경에서도 강건성을 유지할 수 있는 탐지 알고리즘 고도화와, 보호 노이즈의 시각적 품질 최적

화를 통해 실사용성을 더욱 강화할 예정이다.

본 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지역지능화혁신인재양성사업(IITP-2024-RS-2022-00156287) 및 인공지능융합혁신인재양성사업(IITP-2023-RS-2023-00256629)의 지원을 받아 수행되었으며, 또한 농림축산식품부의 재원으로 농림식품기술기획평가원의 농식품과학기술융합형연구인력양성사업(RS-2024-00397026)의 지원을 받아 수행되었음.

참고문헌

[1] Tan, Chuangchuang, et al. "Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[2] Shan, Shawn, et al. "Fawkes: Protecting privacy against unauthorized deep learning models." 29th USENIX security symposium (USENIX Security 20). 2020.