딥페이크 이미지 판별을 위한 Xception 기반 웹 서비스 개발

박지예[1], 전하경[1], 추민서[2], 김지선[3]
[1]덕성여자대학교 컴퓨터공학전공 학부생
[2]덕성여자대학교 IT미디어공학전공 학부생
[3]아마존 웹서비스 코리아
jiye0710@duksung.ac.kr, jeonfuture0120@duksung.ac.kr, 20200925@duksung.ac.kr,
jisunaudreykim@gmail.com

Development of A Web Service Based on Xception for Deepfake Image Detection

Jiye Park[1], Hakyung Jeon[1], Minseo Choo[2], Jisun Kim[3]
[1]Dept. of Computer Engineering, Duksung Women's University
[2]Dept. of IT Media Engineering, Duksung Women's University
[3]Amazon Web Services Korea

요 약

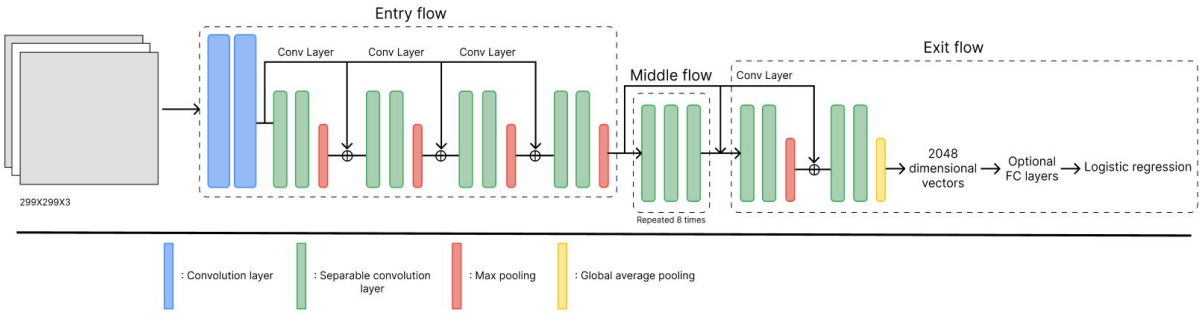
최근, 해외뿐만 아니라 국내에서도 딥페이크 기술을 악용한 피해 사례는 가파르게 증가하는 추세를 보이고 있다. 딥페이크 기술은 엔터테인먼트, 예술, 교육 등에서 창의적 사용이 가능하지만, 동시에 법적, 사회적, 윤리적으로 큰 우려를 일으키고 있다. 본 논문에서 제안하는 딥페이크 이미지 판별 웹 서비스는 딥페이크 기술을 이용한 위조 이미지를 탐지할 수 있어, 사전적 예방 효과를 기대할 수 있다. Xception 모델을 기반으로 한국인 KoDF 데이터 셋을 학습시켜 다양한 위조 이미지 감지 성능을 평가하였으며, 이를 웹 애플리케이션 형태로 구현하여 사용성 높은 서비스를 제공할 수 있다.

1. 서론

인공지능의 발전으로 딥페이크(Deepfake)를 이용한 위조 콘텐츠 제작이 점점 정교해지고 있다. 온라인 상에 존재하는 딥페이크 영상의 96%가 불법 음란 동영상으로 확인되는 사례로 미루어 볼 때, 이는 심각한 사회적 문제로 이어질 가능성이 매우 크다.[1] 이를 방지하기 위해, 다양한 딥페이크 탐지 서비스가 개발되고 있지만 대부분의 서양인 데이터 셋 기준으로 최적화되어 있어 한국인에 대한 위조 피해를 예방하기엔 한계가 있다. 이러한 점에 입각하여, 한국인 데이터셋으로 학습시킨 Xception 모델 기반의 딥페이크 탐지 웹 서비스를 개발하였다.

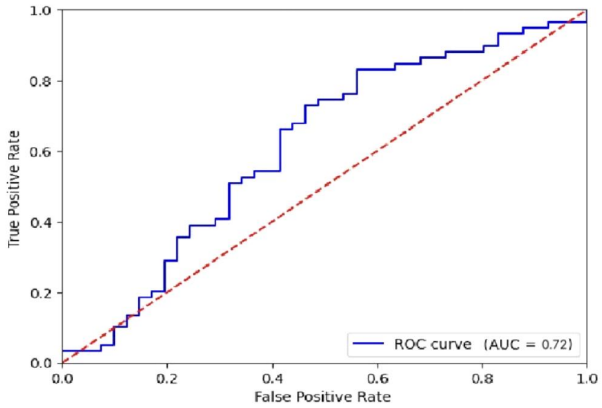
2. 딥페이크 이미지 판별 모델

모델의 학습을 위해 인코더로 Xception 모델을 사용하였다. Xception 모델은 Depthwise separable convolution이 기반이 되며 일반적인 Convolution layer와 비교했을 때 매개변수와 계산 비용을 크게 줄일 수 있는 장점이 있다. 해당 구조는 그림 1과 같이 계산 측면에서 효율적일 뿐만 아니라 딥페이크와 같은 이미지의 복잡한 패턴을 포착하는 데 유리하다.[2] 모델 학습에 사용된 데이터셋은 한국인 KoDF 영상에서 real 이미지와 fake 이미지로 나누어 사용했다.[3] 총 1,985개의 데이터 중 1,588개의 데이터를 학습시킨 후, 나머지 397개의 데이터로 모



(그림 1) Xception 모델 아키텍처

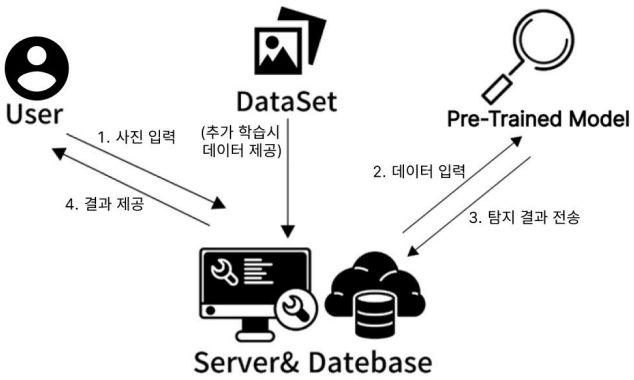
델의 평가를 진행했다. 그림 2는 한국인 데이터셋 KoDF를 학습시킨 Xception 모델의 ROC Curve이다. 모델의 정확도는 전체 데이터셋 중 예측 결과와 실제 값이 같은 비율을 측정하기 위해 (TN + TP) / (TN + FP + FN + TP)로 계산한다. 모델을 훈련한 후 테스트한 결과, 72%의 정확도를 보였다.



(그림 2) 훈련된 모델의 ROC Curve

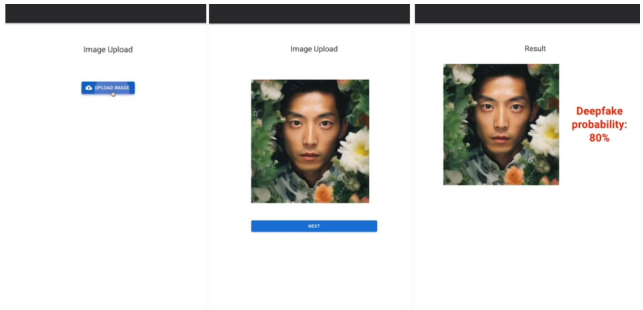
3. 딥페이크 이미지 판별 웹 서비스 개발

업로드된 사용자 제공 이미지를 통해 딥페이크 이미지를 손쉽게 판별하는 웹 서비스를 개발하였다. 웹 구현 시 사용자 호환성을 위해 React를 이용하여 반응형 웹으로 구현하였으며, 웹 서비스 흐름도는 그림 3과 같다.



(그림 3) 서비스 흐름도

그림 4는 Xception 모델을 이용하여 구현한 웹 서비스 화면이다. 사용자의 이미지에 대한 딥페이크 판별 결과를 확인할 수 있다.



(그림 4) 판별 진행 과정 및 테스트 결과 화면

4. 결론 및 향후 연구

본 연구에서는 딥페이크 이미지의 조기 탐지 및 예방을 위해 반응형 웹 기반의 접근성 편한 서비스를 개발하였다. 특히, 기존의 서양인을 대상으로 한 딥페이크 탐지 모델의 한계를 극복하고자 한국인 데이터셋을 적극적으로 사용하였다.

다른 딥페이크 탐지 모델들과 비교했을 때, 모델의 정확도는 다소 낮은 편으로, 이는 학습에 사용한 데이터 셋의 규모가 상대적으로 작아 발생한 문제일 것으로 예상된다. 향후, 양질의 데이터셋을 확보하고 해당 모델의 추가적인 학습을 진행하여 딥페이크 감지 서비스를 견고하게 보강할 예정이다.

참고문헌

[1] Ivan Mehta, A new study says nearly 96% of deepfake videos are porn [Internet], https://thenextweb.com/news/a-new-study-says-nearly-96-of-deepfake-videos-are-porn

[2] Francois Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions”, Computer Vision and Pattern Recognition (CVPR), pp.1251-1258, 2017.

[3] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, Gyeongsu Chae, “KoDF: A Large-scale Korean DeepFake Detection Dataset”, 2021 International Conference on Computer Vision (ICCV), 2021.

-본 논문은 과학기술정보통신부 대학디지털교육역량강화사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.-