Mesoscopic Network를 이용한 딥페이크 감지 기법

이혜리*, 양희규**, 추현승*
*성균관대학교 전자전기컴퓨터공학과
**성균관대학교 수퍼인텔리전스학과
2hyel05@g.skku.edu, huigyu@g.skku.edu, choo@skku.edu

Deepfake Detection with Mesoscopic Network

Hyeri Lee*, Huigyu Yang**, Hyunseung Choo*
*Dept. of Electrical and Computer Engineering, Sungkyunkwan University
**Dept. of Superintelligence Engineering, Sungkyunkwan University

요 약

소셜 미디어와 스마트폰의 대중화로 인해 디지털 이미지와 비디오를 만들어 내는 일이 매우 흔해졌다. 전통적인 이미지 포렌식 기술 압축 방법은 데이터를 손상시킨다는 점에서 비디오에 적용하기 부적절하다. 따라서 본 논문에서는 딥러닝과 MesoNet을 이용한 모델을 통해 참 혹은 거짓만 나타내는 기존의 결과 산출 방법에서 더 나아가 네가지의 분류 방법으로 딥페이크 감지 흐름을 살펴보고자 한다.

1. 서론

미디어 콘텐츠의 다양화는 새로운 형태의 이미지 혹은 비디오 콘텐츠의 재생산으로 이어진다. 최근 통계 결과에서 하루 동안 유통되는 동영상과 이미지 콘텐츠의 양은 약 20억개이며, 가짜 뉴스 혹은 비디오의 디지털 위조 및 변조가 확산되는 추세이다[1]. 대부분의 디지털 콘텐츠 위조 및 변조는 특정 인물의 얼굴이나 신체를 다른 사람의 것으로 합성하는 딥페이크(e.g. Deep Learning과 Fake의 합성어) 형태로 발생한다. 심층신경망 학습 기반의 콘텐츠 생성은 기존 디지털 변조 방식(i.e. 포토샵)과 달리 결과물의 실제 여부를 구분하기 어렵고 재생산이 용이하기 때문에 무분별한 악의적 콘텐츠 생성 문제가 발생한다. 최근에는 딥페이크를 활용한 가짜 뉴스, 금융사기, 리벤지 포르노, 개인 정보 침해, 불법적인 신원 확인 등이 사회 안전을 위협하는 요소로 대두되고 있다. 따라서 딥페이크 기술의 불법적 악용을 방지하기 위한 감지 기술이 필요하다.

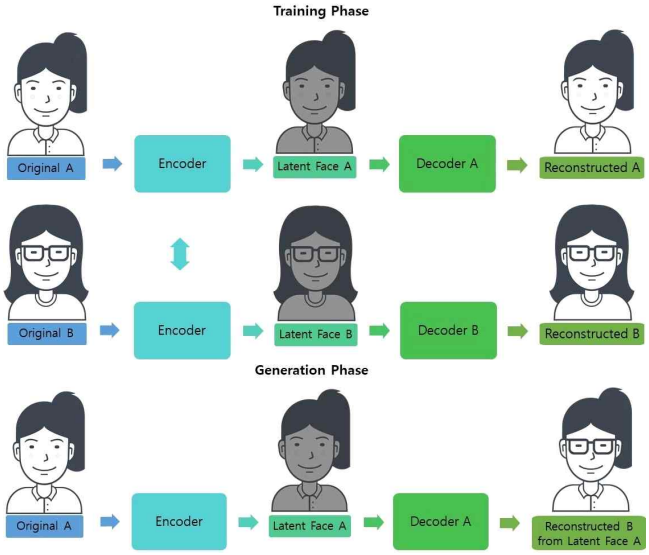
딥페이크 콘텐츠는 입력 이미지를 학습된 심층신경망의 가중치에 의해 변형시키거나 근접한 이미지로 재생성한다. 딥페이크 감지 기술은 생성과정에서 콘텐츠에 포함된 노이즈 또는 텍스쳐의 부자연스러움 등을 픽셀 간의 상관관계를 통해 분석하여 딥페이크 영상을 판별한다. Microscopic 레벨의 분석 기법은 이미지 노이즈를 기반으로 딥페이크 영상을 판별하기 때문에 노이즈의 타격을 많이 받는 압축 비디오에서는 적용될 수 없다. Macroscopic 레벨의 분석은 거시적 접근이 가능한 방법으로 인간의 인지적 특성을 모방하여 눈, 코, 입의 모양이나 피부의 텍스처 등의 부자연스러움을 판별한다. 하지만 변조된 영상이 사람의 눈으로 판별이 어려운 경우 그 신뢰도가 저하된다. 이러한 문제점 해결을 위해서는 Microscopic과 Macroscopic 레벨 분석 방식의 중간 단계인 Mesoscopic 레벨의 접근 방식이 필요하다. 따라서 본 논문에서는 픽셀 단계의 미시적인 정보와 인지적 특성을 포함한 거시적인 정보를 동시에 학습할 수 있는 MesoNet(Mesoscopic Network) 기반의 감지 모델을 제안한다.

2. 관련연구

2.1 딥페이크 생성

딥페이크는 FakeApp이라는 어플리케이션으로 2018년에 처음 등장하였다[2]. 딥페이크 생성은 크게 두가지 방법으로 나뉘어지는데, GAN(Generative Adversarial Network)을 이용한 방법과 그림 1과 같은 autoencoder를 이용한 방법이다. 딥페이크 비디오 생성은 크게 추출, 학습, 생성의 세가지 단계를 거친다[3]. 추출 과정에서는 데이터로부터 모든 프레임들이 추출된 뒤, 얼굴을 식별하고 정렬한다. 학습 과정에서는 가중치들이 최적화되는 단계이다. 이미지가 encoder 입력값으로 주어지면 숨어있던 latent

face가 만들어지고 decoder에 들어간 후에는 재구성된 입력 이미지가 탄생한다. 딥페이크 생성을 위해서는 원본 얼굴과 target을 위한 두 세트의 autoencoder와 decoder가 필요하다. 학습 과정에서, 두 encoder 모두 공통된 특징점을 위해 가중치를 공유하고 그 특징점들을 학습한다. 학습 과정이 끝난 후 생성 단계에서 latent vector A가 인공지능 모델을 통해 압축되고 동일한 사람에 대한 정보를 재생성하기 위해 입력 이미지 A를 사용한다.



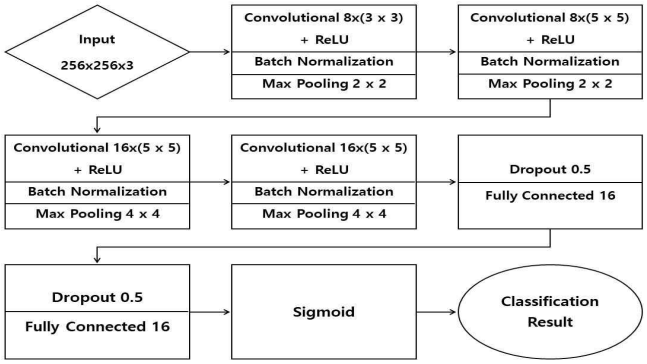
(그림 1) Autoencoder를 이용한 딥페이크 생성

2.2 딥페이크 감지

앞서 언급된 여러 가지 사회적 이슈로 인해, 최근 딥페이크 감지에 대한 연구가 활발하게 진행되고 있다. 기존의 연구 방법으로는 전통적인 방식의 연구와 심층신경망을 이용한 연구가 있다. 전통적인 방식의 연구는 딥페이크 감지를 위해 이미지나 비디오의 픽셀 정도 차이를 체크한다. 픽셀 간의 상관관계를 통해 진짜 이미지와 가짜 이미지 사이의 변이를 알아내는 일은 복잡하지 않은 과정에 속하나, 이미지가 변형될 경우 robustness를 고려해야 한다는 문제가 있다[4]. 심층신경망을 이용한 연구는 공간적 특성 분석, 감지 효율성 증대, capacity 증가, 생성 단계에서 발생하는 시공간적 결함 해소를 목적으로 하고 있으나 적대적 공격(adversarial attack)에 취약하다는 특징이 있다[5].

2.3 Meso4

Meso4는 딥러닝 기반의 CNN(Convolutional Neural Network)기법을 사용하여 이미지 데이터를 예측하고 매우 사실적으로 조작된 비디오 판별을 위해 제안된 모델이다[1]. Meso4는 mesoscopic의 레벨에서 분석하는 모델로 이는 microscopic과 macroscopic의 중간 단계를 말한다. 딥페이크의 경우 사람의 얼굴을 매우 상세하게 묘사하므로, 육안으로는 조작된 이미지의 구별이 어렵다[6]. 그림 2는 Meso4의 네트워크 구조이다. 기존의 모델들보다 레이어의 개수가 낮고 비교적 간단한 모델이긴 하나, 이는 이미지의 mesoscopic한 특징에 집중하기 위함이다.



(그림 2) Meso4의 네트워크 구조

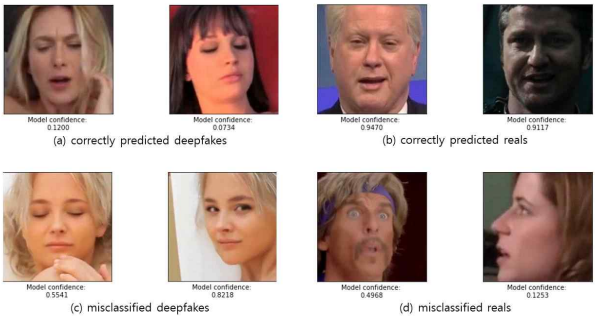
3. 제안모델

본 연구에서는 Meso4모델을 벤치마크로 사용하였다. 일반적인 real 혹은 fake의 두가지 결과만 내는 방법과 달리, correctly predicted deepfakes, correctly predicted reals, misclassified deepfakes, misclassified reals의 네가지 카테고리로 예측 결과를 분류한다. 제안하는 모델의 흐름은 다음과 같다. 픽셀값을 0과 1사이로 정규화하고, 이미지의 범위를 재조정해주는 image data generator를 생성해 데이터 폴더의 직접적인 경로를 구체화한다. Batch size를 1로 두어 이미지에 각각 접근할 수 있게 하고, Binary로 클래스 노드를 설정하여 이미지를 fake 혹은 real로 나올 수 있게 만든다. 다음 단계에서는 변수 x와 y를 generator.next와 같게 설정한 뒤 예측 결과를 평가하게 되는데, 결과 값은 네자리 수이며 예측 값의 반올림이 실제 라벨링 된 결과와 맞는지 확인한다.

4. 실험 및 결과

사용한 데이터셋은 MesoNet의 기존 딥페이크 비디오에서 추출되었으며, 해당 데이터셋은 TV나

유튜브를 통해 수집되었다. 본 모델에서 fake의 결과값은 0, real의 결과값은 1을 나타낸다. 따라서, 예측된 결과값이 0이나 1에 가깝다면 예측의 신뢰도가 높은 것이고, 0.5의 근사치는 모델 예측의 신뢰도가 random guess임을 나타낸다. 그림 3의 (a)와 (b)는 정확한 분류가 이루어졌으므로 model confidence가 0과 1에 매우 가까움을 알 수 있다. 반면에 (c)와 (d)는 오분류의 결과를 나타내며 각각 왼쪽 사진의 confidence는 0.5에 가까우므로 random guess라 판단하였다. (c)는 fake를 real로, (d)는 real을 fake로 분류하였으므로 예측(오른쪽)의 confidence는 각각 1과 0에 가깝다.



(그림 3) Model Confidence

5. 결론 및 향후 연구계획

본 논문에서는 딥페이크 감지 결과를 혼동행렬을 기반으로 네가지로 분류하였다. 특히 Deepfake 영상에 대한 감지 성능이 Real 영상에 대한 감지 성능보다 중요하기 때문에 이를 분리하여 측정하였다. 영상의 레이블과 model confidence를 비교한 결과, Real과 Fake 영상에 대한 평균 오차율은 각각 0.16, 0.19로 나타났다. 향후 연구로는 confidence 향상을 위해 MesoInception-4에 제안모델을 적용하고자 한다. 또한 딥페이크 감지뿐만이 아니라 딥페이크 생성 모델 구현을 통해 높은 감지 정확도를 지니는 GAN 모델을 개발하고자 한다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT명품인재양성 사업 (IITP-2022-2020-0-01821), 지역지능화혁신인재양성 (Grand ICT연구센터) 사업 (IITP-2022-2015-0-00742), 인공지능혁신허브 사업(IITP-2022-0-02068), 한국연구재단의 지원 (NRF-2020R1A2C2008447)을 받아 수행된 연구임.

참고문헌

[1] Afchar, Darius & Nozick, Vincent &Yamagishi, Junichi & Echizen, I. MesoNet: a Compact Facial Video Forgery Detection Network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1-7.

[2] FakeApp. https://www.fakeapp.com/. Accessed:2018-09-01

[3] Mitra, Alakananda & Mohanty, Saraju & Corcoran, Peter & Kougianos, Elias. A Machine Learning Based Approach for Deepfake Detection in Social Media Through Key Video Frame Extraction. SN Computer Science. 2. 10.1007/s42979-021-00495-x. 2021 (2021).

[4] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2019 (pp. 1-11).

[5] Malik, A., Kuribayashi, M., Abdullahi, S.M. and Khan, A.N., DeepFake Detection for Human Face Images and Videos: A Survey. IEEE Access, 10, pp.18757-18775. 2022

[6] V. Schetinger, M. M. Oliveira, R. da Silva, and T. J. Carvalho. Humans are easily fooled by digital images. arXiv preprint arXiv: 1509.05301, 2015. 3