Two-Stream 어텐션 기반의 Efficientnet과 Bi-LSTM을 이용한 딥페이크 영상 탐지

연세대학교 대학원

산업공학과

송 나 연

Two-Stream 어텐션 기반의 Efficientnet과 Bi-LSTM을 이용한 딥페이크 영상 탐지

지도 김 우 주 교수

이 논문을 석사 학위논문으로 제출함

2020년 12 월 24 일

연세대학교 대학원

산업공학과

송 나 연

송나연의 석사 학위논문을 인준함

심사위원 김 우 주 인

심사위원 김 창 욱 인

심사위원 홍 준 석 인

연세대학교 대학원

2020년 12 월 24일

차 례

< 그림 차례 >

< 표 차례 >

국 문 요 약

Two-Stream 어텐션 기반의 Efficientnet과 Bi-LSTM을 이용한 딥페이크 영상 탐지

연세대학교 일반대학원

산업공학 전공

송나연

현재 딥러닝 기술은 여러 분야에서 성공적으로 적용되어 이용되고 있다. 특히 컴퓨터 비전 분야에 있어 딥러닝 기술의 발전을 통해 많은 진보를 이루어냈다. 하지만 기술이 발전됨에 따라 개인 프라이버시, 국가 안보 등에 위협을 주는 기술의 악용된 사례들도 비례하여 나타나고있다. 이러한 기술들 중 하나가 바로 딥페이크 기술이다. 딥페이크 기술은 산업적으로 활용 가치가 높기 때문에 영화, 뮤직비디오 등 다양한 분야에서 활용되기도 하지만[38], 동영상 조작을 통해 진실을 왜곡하여 악용되고 있는 문제점이 제시되었다. 딥페이크의 악용에 대한 우려가 커짐에 따라 동영상의 조작 여부를 판단하는 연구가 활발히 진행되고 있다.

본 논문에서는 합성곱 신경망(Convolutional neural network ,CNNs) 와 순환

신경망(Recurrent neural networks , RNNs) 를 사용하여 프레임 속 시각적 정보와 프레임 간 시간적 정보를 추출하여 딥페이크 조작 여부를 정확하게 감지하고자 한다. 기존 연구들에서는 합성곱 신경망를 이용하여 시각적 정보만 이용하여 프레임 간 시간적 정보를 포함하지 못한다는 한계가 있다. 뿐만 아니라 최근 연구들은 프레임 속 얼굴 정보에 한정되어 딥페이크 영상을 탐지해 조작 여부를 조작이 발생한 부분에 한정되어 탐지한다는 한계가 존재한다.

본 연구에서는 동영상 내 전역 정보(Global Information)과 지역 정보(Local Information), 두 가지 정보에서 특징을 추출함으로써 보다 효과적인 딥페이크 영상을 탐지하고자 한다. 특히 지역 정보는 조작이 일어나는 영상 속 얼굴 정보를 의미하고, 글로벌 정보는 영상 전체의 조화를 판단할 수 있는 프레임 전체 정보를 의미한다. 따라서 이 두 가지 정보를 융합하여 두 정보의 상보성을 보완한 딥페이크 영상 탐지 방법을 제안한다. 본 논문은 FaceForencies++ 데이터 셋을 사용하여 평가하고 다른 모델에 비해 경쟁력 있는 결과를 제공한다.

Keyword: 조작된 비디오 탐지, 딥페이크 탐지, Efficientnet, 비디오 포렌식 탐지, Two-Stream

1. 서론

인류의 역사에서 가장 우위에 있는 인간의 자각 능력을 꼽으면 주로 ‘시각’을 말할 수 있다. 동 서양을 막론하고 ‘백문이 불여일견’과 ‘Seeing is believing’과 같은 속담들이 통용되어 사용되는 이유도 시각 인지에 대한 확고함 때문이라고 할 수 있을 것이다. 눈으로 보는 이미지는 가장 강력한 진실의 척도였고, 어떤 판단에 있어 시각에 대한 의지는 불가침의 영역이었다. 하지만 딥러닝으로 대표되는 디지털 기술의 발전으로 인해 시각 인지에 대한 인류의 믿음을 송두리째 흔들고 있으며 그 정점에 서 있는 기술이 최근 들어 사회문제로도 대두되고 있는 ‘딥페이크(Deepfake)’기술이다.

딥페이크는 인공지능을 기반으로 원본 이미지 위에 다른 이미지를 중첩하거나 결합시켜 원본과 쉽게 구분할 수 없는 가공의 이미지와 소리를 갖도록 만들어지는 영상 혹은 그 제작과정을 말한다[1]. 인터넷, 소셜미디어에서 실제로 촬영된 비디오가 아닌 인공지능을 활용하여 제작된 콘텐츠, 딥페이크(Deepfake)가 빠르게 확산되며 사회적 관심이 커지고 있다. 딥페이크(Deepfake)는 딥러닝(Deep Learning)을 이용하여 원본 동영상 위에 다른 동영상을 중첩하거나 결합하여 새로 조작된 콘텐츠를 생성하는[38] 기술이다. 딥페이크 알고리즘이 공개 됨으로 인해, 가짜 이미지나 합성사진을 생성하기 위해 포토샵과 같은 이미지 편집 프로그램을 사용해야했던 과거와 달리 손쉽게

딥페이크로 조작된 이미지와 동영상을 생성할 수 있다. 2017년 미국의 레딧 사이트에 유명 배우의 얼굴로 조작된 가짜 포르노 비디오가 올라오는걸[2] 시작으로 이 기술을 이용해 유명인사들의 얼굴을 바꿔서 만든 가짜 영상이 우후죽순 등장했다. 2019년 9월 기준 딥페이크 영상 중 96%는 포르노이며, 총 영상 조회수는 1억이 넘는 등 문제가 되고 있다[3]. 뿐만 아니라, 마테로 렌치(Matteo Renzi) 전 이탈리아 총리가 다른 정치인을 모욕하는 딥페이크 영상[34]으로 인해 사회적 파문을 일어났고, 멕시코에서는 딥페이크 동영상을 통해 대통령 후보를 비하는 등 딥페이크를 활용한 가짜 영상으로 정치적, 사회적 불안감이 가중되고 있다.

이처럼 딥페이크 기술의 악용이 커다란 사회적 파장을 일으킬 수 있는 만큼, 딥페이크 영상 탐지에 대한 연구도 활발히 진행되고 있다. 프레임 전체가 조작되고, 이를 탐지하는 과거의 조작 영상 탐지 연구와 달리, 오늘날의 딥페이크 영상은 영상 속 인물의 얼굴에 국한되어 조작이 진행되고, 이를 탐지한다. 진행중인 딥페이크 영상 탐지 연구는 크게 영상에서의 비정상적인 움직임에 기반하는 기술[4], 이미지 자체의 품질에 기반하는 기술[5], 그리고 딥러닝에 기반하여 탐지하는 기술[6]로 나눌 수 있다. 얼굴의 비정상적인 움직임을 탐지하는 기술은 딥페이크 기술의 지속적인 발전에 대응하지 못하고, 이 탐지 방법은 학습을 통해서 충분히 피해갈 수 있는 문제점이 있다. 이미지 품질을 측정하여 탐지하는 기술은 딥페이크 기술이 많은 학습량을 통해 고해상

도의 결과를 도출하면 탐지가 제한되는 문제점이 존재한다. 따라서 본 연구를 프레임 속 시각적 정보(Visula Information)과 시간적 정보를 모두 활용하는 딥페이크 영상 탐지를 하고자 한다. 일반적으로 딥페이크 기술은 얼굴에 국한되어 발생한다. 이에 따라 대부분의 연구들이 프레임 속 얼굴을 탐지한 후, 얼굴 정보를 이용하여 영상의 조작 여부를 판별해 낸다[7][8][9]. 본 연구에서는 동영상 내 전역 정보(Global Information)과 지역 정보(Local Information), 두 가지 정보에서 특징을 추출함으로써 보다 효과적인 딥페이크 영상을 탐지하고자 한다. 특히 지역 정보는 조작이 일어나는 영상 속 얼굴 정보를 의미하고, 글로벌 정보는 영상 전체의 조화를 판단할 수 있는 프레임 전체 정보를 의미한다. 이 두 가지 정보를 융합하여 두 정보의 상보성을 보완한 딥페이크 영상 탐지 방법을 제안한다. 또한, FaceForencies++ 데이터[10]를 사용하여 평가하여 다른 모델에 비해 경쟁력 있는 결과를 제공한다.

본 논문은 다음 순서로 구성되었다. 2장에는 딥페이크 조작 방식, 딥페이크 탐지를 위한 얼굴 추출 모델 그리고 딥페이크 연구의 동향을 소개한다. 3장에서는 본 논문에서 제안하는 CNN과 RNN기반의 모델과 Attention mechanism이 각 모델에 어떻게 적용되어 있는지 상세하게 설명한다. 4장에서는 사용된 데이터와 각 모듈에서 다양한 모델들을 비교 실험함으로써, 제안한 모델의 성능을 평가하고 이를 분석한다. 마지막으로 5장에서는 본 연구의 결론에 대하여 서술한다.

2. 관련 연구

2.1 얼굴 조작 (Face Manipulation)

딥페이크 기술, 즉 영상을 조작하는 방법에는 다양한 방법론들이 존재하지만 일반적으로 사용되는 기술은 Deepfakes[11], Face2Face[14], FaceSwap[13], NeuralTextures [12] 4가지이다. 이 조작법은 크게 컴퓨터 그래픽 기반 방법론과 학습 기반 방법론으로 나눌 수 있고, 변경 된 부분에 따라 신원 조작(Identity Manipulation)과 표정 조작(Expression Manipulation)으로 나뉜다. 최근 이 기술들을 이용한 비디오 기반 얼굴 조작 데이터들이 공개됨과 동시에 이용 가능해졌다. FaceForensics는 Face2Face 기술을 이용한 조작된 비디오 데이터를 공개했고, 이후에 Deepfake 기술과 Face Swap 기술을 통해 조작된 비디오 데이터를 증강한 FaceForensics++[10]라는 발전된 데이터까지 공개했다.

Face2Face와 FaceSwap는 컴퓨터 그래픽 기반 방법이다. Face2Face[13]는 소스 영상에서 선택된 얼굴의 표정이 변경하고자 하는 영상의 얼굴로 전달한다. FaceSwap[14]은 얼굴 랜드마크 정보를 사용하여 소스 영상의 얼굴을 변경하고자 하는 영상으로 투입하고, 랜드마크 간의 차이를 최소화 시킨 후 색상

을 정밀하게 변경함으로서 실제와 같은 영상을 생성해 낸다. Deepfakes와 Neural Textures는 학습 기반 조작방법으로써 Deepfakes[11]는 두개의 오토인코더(AutoEncoder)가 각기 다른 두개의 영상의 얼굴을 재구성하도록 학습한다. 영상을 인코더(Encoder)로 압축시키고, 다른 영상의 얼굴을 재구성하도록 만든 디코더(Decoder)를 활용하여 원본이미지를 다른 영상으로 복원시켜 영상을 조작한다. Neural Textures[12]는 변경하고자 하는 영상의 뉴럴 텍스쳐를 학습하여 표정을 변경함으로써 Face2Face보다 자연스럽게 영상을 조작한다. 이 외에도 최근에는 생산적 적대적 신경망(GAN), 합성곱 신경망 등 다양한 딥러닝을 기반으로 영상을 조작하는 기술에 대한 연구가 진행되고 있다.

2.2 딥페이크 (Deepfake)

일반적으로 딥페이크 데이터는 앞서 언급했듯이 Deepfakes[11], Face2Face[14], FaceSwap[13], NeuralTextures [12] 4가지 방식을 통해 조작된 영상을 의미한다. 그 중 Deepfake[11] 기술은 학습을 기반한 딥러닝 모델을 이용한 동영상 조작 방법으로 크게 face extraction, learning, generation의 단계로 이루어져있다. Face extraction 단계에서는 바뀌는 얼굴 이미지의 특징을 우선 확보하는 단계로 학습하는 모델이 얼굴이 가지고 있는 특징을 학습한다. 입력 이미지에 대해 얼굴을 탐지하고 추출한 후, 추출한 얼

굴을 위치에 맞게 정렬하여 마무리 한다. 다음 단계인 learning 부분에서는 앞서 추출한 얼굴의 특징을 조작하고 싶은 얼굴로 다시 생성하도록 하는 과정이다. 입력 이미지를 인코더(Encoder) 부분에서 압축시키고, 디코더(Decoder) 부분에서 다시 출력이 동일하도록 생성하는 오토인코더를 이용하여 딥페이크 학습이 이루어진다. 학습 과정이 끝나면 재구축된 얼굴을 원본 이미지에 병합하여 생성하는 단계로 넘어간다. 현재 딥페이크 기술은 엔비디아가 개발한 StyleGAN[15]을 이용하여 완전히 새로운 가짜 얼굴 이미지를 만들어 내는 기법 등으로 지속적인 발전을 보이고 있다.

2.3 얼굴 탐지 (Face Detector)

얼굴 탐지 기술은 비디오 내에 존재하는 얼굴을 검증(verification)과 식별(identification)하는 기술로 구분되어 발전되어 왔다. 이후 딥러닝의 발전으로 인해 컴퓨터 비전 분야에서 압도적인 성능을 보였다[16]. 이에 따라 CNN(Convolutional neural networks)기반의 딥러닝 모델이 전통적인 방식을 대체하게 되었다. 얼굴을 탐지하기 위해서는 얼굴의 영역을 정확히 검출 되어야 하는데 CNN기반의 대표적인 모델이 MTCNN[17]이다.

MTCNN[17]은 얼굴 영역을 검출해 내는 Face detection과 눈, 코, 입의 위치를 검출하는 Face Alignment, 이 두가지 영역이 서로 긴밀한 연결이 있다는

가정하에 얼굴 위치의 세부 조절을 해주는 bounding box regression, 총 세 가지 테스크를 동시에 학습시킨다. 결과적으로 세 가지 테스크를 각각 학습한 P-net, R-net, O-net의 CNN을 차례대로 통과하는 Cascade 모델을 제안한다. 첫 번째 단계인 P-net에서는 작은 이미지를 입력으로 받아 컨볼루션 레이어를 거쳐 얼굴 영역을 검출해내는 x, y, w, h로 이루어진 bounding box regression 결과값, 검출 해낸 영역이 얼굴인지 아닌지 나타내는 face classification 결과값, 그리고 양쪽 눈, 코, 입의 좌표를 나타내는 10개의 landmark localization 결과값을 다음 단계로 전달한다. P-net은 아주 작은 크기의 윈도우로 작은 얼굴도 찾아낼 수 있다는 장점이 있다. 이후 R-net에서는 P-net을 통해 얼굴로 추정되는 박스의 리스트를 가지고 크기를 키워주는 작업을 한 후에 이 박스들 중에 진짜 얼굴에 해당하는 영역들을 추려내고, bounding box regression을 더욱 정교하게 수행한다. R-net에서 찾아낸 박스는 O-net으로 전달된다. O-net 에서도 R-net을 통해 찾아낸 박스들의 크기를 키워준다. 이렇게 이미지 필터의 크기를 키우면서 얼굴에 해당하는 더욱 추상적인 정보를 찾아낼 수 있게 된다. 이와 같이 MTCNN은 이미지 피라미드를 이용하여 다양한 크기의 이미지를 3가지의 네트워크로 학습시킨 모델로 실시간 처리가 가능하며 하나의 이미지에서 다양한 얼굴의 크기를 검출할 수 있다.

얼굴 탐지(face detection)기술은 본 연구에서 사용하는 MTCNN이외에도 BlazeFace[18] 등 다양한 기법들이 연구되어 발전을 거듭하고 있는데, 이 모

델은 전체 이미지 또는 비디오 프레임에 대하여 2D, 3D로 얼굴의 키 포인트, 윤곽선, 표면 형상 등을 추정하여 적절한 얼굴의 관점에서 정의되는 영역을 탐지하여 얼굴을 검출할 수 있다.

2.4. 딥페이크 탐지 (Deepfake Detection)

딥페이크 기술이 발전됨과 동시에 악용하는 사례들이 늘어남에 따라 조작된 영상을 탐지하기 위해 많은 기술들이 개발되고 있다. 딥페이크를 탐지하는 기법은 크게 세가지로 볼 수 있다. 가짜 영상에서 발생할 수 있는 비정상적인 움직임을 기반으로 탐지하는 기술, 이미지를 기반으로 하는 탐지 기술, 그리고 인공지능 기술을 통해 탐지하는 기술이 존재한다. 첫 번째 방법은 딥페이크로 조작된 영상에서 나타나는 비정상적인 눈 깜박임에 집중하여 탐지하는 기술[4], 비디오와 오디오의 특징을 동시에 추출하고, 입술의 움직임과 목소리의 패턴이 일치하는지를 탐지하는 기술[35]등이 존재한다.

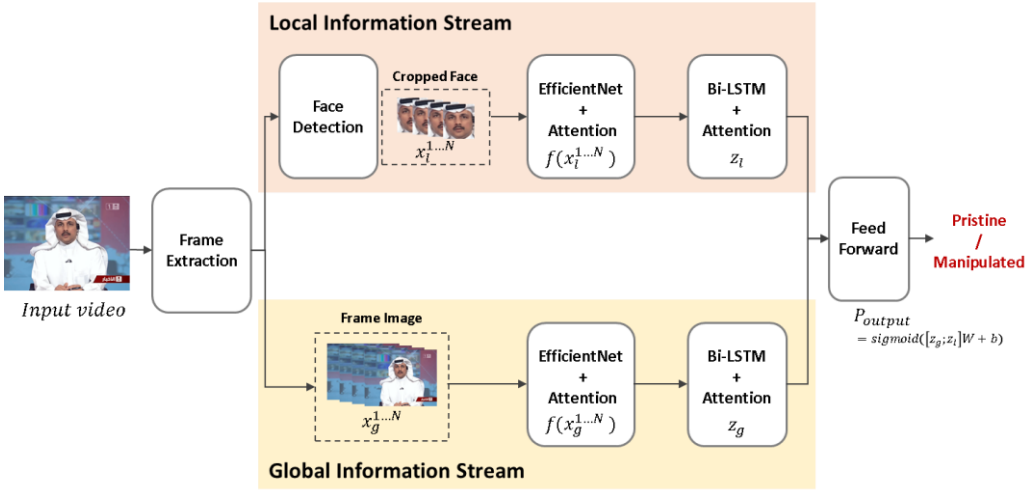
두 번째 방법은 이미지 자체의 품질을 기반으로 원래 얼굴과 조작된 얼굴의 특징을 PCA-LDA(Principal Component Analysis-Linear Discriminant Analysis) 분류기를 가지고 탐지하는 방법[19], 압축 등의 문제로 떨어지는 이미지의 품질을 측정하여 탐지하는 방법[20]등의 방법이 존재한다.

하지만 앞서 두 방법들은 딥페이크 기술의 발전으로 인해 고품질의 가짜 영

상이 나오는 현재 상황의 문제점에 적합하지 않다는 한계가 존재한다. 이를 극복하기 위해 인공지능을 적용하여 탐지하는 기법들이 개발되어 가고 있다. CNN을 이용하여 영상으로부터 프레임 별 이미지를 추출하여 SVM(Support Vector Machine)을 이용하여 분류하는 방법, RNN계열의 LSTM을 이용하여 프레임이 딥페이크인지 오리지널 영상인지 구분하는 방법[21]등이 개발되고 있다.

3. 연구 방안

본 연구에서는 우선 프레임 내에서 특징을 추출한 후, 프레임 간 특징을 추출함으로써 프레임 내 시각적 정보와 프레임 간 시간적 정보를 모두 담고 있는 특징 벡터(Feature vector)를 추출한다. 기존 연구들이 프레임 내 시각적 정보를 담은 특징 벡터를 추출하기 위해 프레임 전체 이미지 정보(Global Information)를 사용하는 방식[21]과 조작이 발생하는 얼굴 정보(Local Information)만 사용하는 방식[22, 23]으로 나뉘지만, 본 연구는 Two-Stream 기법을 통해 두 가지 정보를 모두 사용한다. 또한 각 단계에서 어텐션 방식(attention mechanism)을 활용하여 딥페이크 탐지에 유익한 특징은 선택적으로 강조하고, 덜 유용한 특징은 억제한 특징 벡터(Feature Vector)를 추출한다. 제안 방법론은 <그림1> 에서 보이듯 전처리 단계, 프레임 단계, 비디오 단계 총 3단계로 나뉜다. 제안 방법론에 대한 자세한 설명은 이어지는 제3.1절 ~제3.3절에서 다루도록 한다.



<그림 1> 모델 전체 흐름도

3.1 전처리 (Preprocess Level)

전처리 단계는 크게 프레임 추출, 얼굴 탐지 및 추출, 이미지 크기 재조정, 정규화로 이루어져 있다. 본 연구에서는 딥페이크 조작 여부를 판별하기 위해 <그림 1>'과 같이 영상을 프레임 단위로 추출한 후, 프레임을 시간순으로 엮어 다시 영상 단위로 조작 여부를 판별한다. 전처리 과정에서는 우선 영상에서 프레임을 추출한다. 추출한 프레임은 두 가지 흐름으로 나뉘어 모델로 전달된다. 전역 정보 스트림으로는 추출한 프레임을 그대로 반영하되, 기존의 프레임 사이즈에서 224*224로 크기만 재조정한다. 지역 정보 스트림은 영상에서 추출한 프레임 속 얼굴만 재추출하여 모델로 전달한다. 탐지한 얼굴은 프

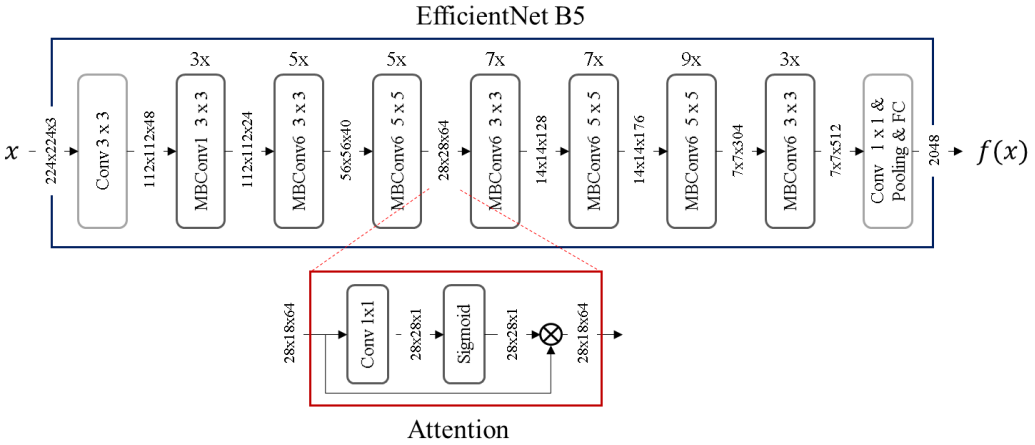
레임별로 크기가 상이함으로, 전역 정보와 동일하게 탐지한 얼굴을 224*224로 크기를 재조정하여 사용한다. 이미지 크기는 이중 선형보간법(Bilinear interpolation)을 통해 2차원의 이미지를 효과적으로 축소한다. 크기가 재조정된 이미지는 정규화를 통해 일정한 범위 내 값으로 변환되어 모델로 입력된다. 본 연구에서는 0에서 255 사이의 이미지 값을 0에서 1 사이의 값으로 스케일링(Scaling)한다.

3.2 시각적 정보 추출 (Frame-Level)

3.2.1 Two-Stream 기법

FaceForensics++[10] 의 데이터는 영상 내 얼굴을 탐지한 후 얼굴 영역 내에서 표정 및 신원을 조작한다. 딥페이크 조작이 발생하는 영역에서는 조작이 발생하지 않은 부분과 대비 차이, 색상의 불규칙성, 흐림 현상 등과 같은 불일치성이 발생한다[24]. 본 연구에서는 두 가지 정보를 활용하여 불일치성을 포착하고자 한다: (i) 조작이 발생하는 부분과 발생하지 않은 부분 간의 불확실성을 포착하기 위해 전역 정보(Global Information) 활용 (ii) 얼굴 영역 내에서 발생하는 불규칙성을 포착하기 위해 지역 정보(Local Information) 활용. 전역 정보 스트림(Global Information Stream)에서는 영상에서 추출한

프레임 속 모든 시각적 정보를 모델에게 전달한다. 반면에 지역 정보 스트림(Local Information Stream)은 조작이 발생하는 얼굴 속 시각적 정보를 모델에게 전달한다. 지역 정보 스트림을 통해 전역 정보에서 포착하지 못했던 조작 발생 지역의 세부적인 불일치성을 포착함으로써 효과적으로 딥페이크 조작을 탐지해내고자 한다.



<그림 2> Efficientnet b5 Attention 모델 구조도

3.2.2 특징 추출 (Feature Extraction)

페이스북의 딥페이크 탐지 대회(Deepfake Detectioni Challenge, DFDC)[25] 와 딥페이크 탐지 분야의 공식 벤치마크[26] 등 프레임 내 특징을

추출하여 딥페이크 여부를 판별해내는 연구가 활발하게 진행되고 있다. 이들 중 높은 정확성을 보이는 모델들은 ILSVRC에서 수상한 CNNs 기반 모델을 활용하여 프레임 내 존재하는 시각적 특징을 추출한다. 본 연구에서는 Image Classification task에서 높은 성능을 보이는 Efficientnet을 통해 전역 정보와 지역 정보 내 중요한 특징을 추출하고자 한다[27]. Efficientnet은 기존의 ImageNet 데이터셋에 대해 정확도와 효율성을 초점으로 제안된 모델들보다 획기적으로 높은 성능과 좋은 효율성을 보여준다. Efficientnet 중에서도 본 연구에서는 모델 크기, 속도, 정확도 간의 균형적인 트레이드 오프를 제공하는 EfficientNet-b5[27]를 사용하였다. FaceForensics++ 데이터는 데이터 수가 조작된 데이터에 편중된(제4.1절) 데이터 불균형 문제가 존재한다. 따라서 ImageNet 데이터 셋 사전 학습된 모델을 사용했고, FaceForensics++ 데이터 셋에 적합하도록 추가 학습했다.

Efficientnet은 <그림 2>와 같이MBConv를 활용하여 구축되어있다. 본 연구에서는 모델 중간에 BAM[28] 에서 제안한 Spatial Attention을 응용한 Attention mechanism을 추가하여 모델의 정확도를 높였다. 본 연구에서 사용된 Attention Mechanism은 다음과 같다.

$$M(F) = \sigma(A \cdot F) \cdot F$$
$$where, \quad F \in R^{C \times H \times W}, \quad A \in R^{1 \times H \times W} \tag{1}$$

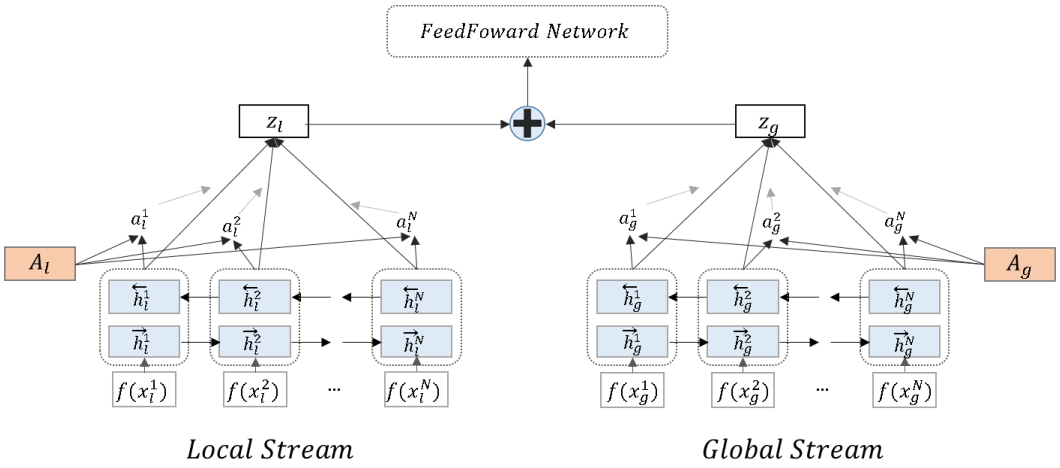
F는 EfficientNet의 중간 Feature map으로써 본 연구에서는 H=28, W=28, C=64이다. A는 F에 1x1 conv 필터를 적용한 결과로써 크기는 H=28, W=28이다. 두 특징을 곱한 후 시그모이드 함수를 통해 Attention Score를 구한 후 다시 F를 곱해 전역 정보와 지역 정보를 모두 반영하기보다는 딥페이크 탐지에 중요한 정보는 선택적으로 강조하고, 불필요한 정보는 억제한 특징 벡터를 추출한다.

또한, 본 연구에서는 비교 모델로 Efficientnet-b4를 사용하여 시각적 특징을 추출하였다. Efficientnet-b4는 Efficientent-b5보다 Imeagenet 데이터에 대해 약간 낮은 성능을 갖고 있다. 하지만 Efficientnet-b5 보다 약 100만 개 적은 파라미터 수를 보유해 빠르고, 가볍다는 특징이 있다[27]. 비슷한 Trade-off를 갖고 있는 Efficientnet-b4와 Efficientnet-b5를 비교 실험을 통해 선택 모델의 우수성을 증명하고자 한다.

3.3 시간적 정보 추출 (Video-Level)

사진과 달리 영상은 시간적 정보를 보유하고 있다. 시간적 정보를 딥페이크 탐지에서 활용한 방식은 다음과 같다. 먼저 영상에서 프레임 정보를 추출한다. 그다음 추출한 프레임 정보에서 딥페이크 탐지에 유용한 특징만 추출한

다. 마지막으로 프레임마다 추출한 정보를 시계열로 묶어 프레임 사이의 시간적 정보를 고려한 특징 벡터를 추출한다. 영상에서 프레임, 다시 영상 형태로 변환시키는 과정을 통해 영상 내의 시각적 정보, 시간적 정보를 모두 고려한다. 본 연구는 시간적 정보를 잘 활용하기 위해 순환 신경망(Recurrent neural networks, RNNs)을 사용했다. 1초당 약 30개의 프레임이 나오는 영상의 정보를 효과적으로 반영하기 위해 LSTM(L Long Short-Term Memory)을 선택했고, 정방향(과거에서 미래 방향)과 역방향(미래에서 과거 방향)을 모두 고려한 양방향 LSTM (Bidrectional LSTM)[29] 을 사용한다.



<그림 2> Bi-LSTM Attention 모델 구조도

3.2부 절에서 진행된, Efficientnet에서 추출된 $f(x_n)$를 양방향 LSTM에 입력하여 프레임 간 시각적 정보를 반영한 h_n을 생성한다.

$$\begin{aligned} \overrightarrow{h^n} &= \overrightarrow{LSTM}(f(x^n)), \qquad n \in [1, N], \\ \overleftarrow{h^n} &= \overleftarrow{LSTM}(f(x^n)), \qquad n \in [N, 1] \\ h^n &= \left[\overrightarrow{h^n}, \overleftarrow{h^n}\right] \end{aligned} \tag{2}$$

다음은 Attention Mechanism[30]을 통해 h_n로부터 가중치가 반영된 특징 벡터 z를 얻는다. 시점별로 추출된 시각적 특징이 상이하기 때문에 이에 따른 가중치를 부여하여 합한 하나의 벡터를 구한다

$$\begin{aligned} g_w^n &= tanh(W_w h^n + b_w),\ w \in [l, g] \\ \alpha_w^n &= \frac{\exp(g_w^n)}{\sum_n(\exp(g_w^n))} \\ z_w &= \sum_n \alpha_w^n h_w^n, \qquad n \in [1, N] \end{aligned} \tag{3}$$

최종적으로 영상 전체 정보를 담고 있는 z_l와 z_g를 연결(concatenate)하여 한 번 더 layer를 통과시킴으로써 전역 정보와 지역 정보, 시각적 정보와 시간적 정보를 모두 활용하여 효과적으로 딥페이크 조작 여부를 파악하는 모델을 구축한다.

4. 실험 및 결과

4.1 실험 데이터

학습에는 이탈리아 나폴리 페데리코 2세 대학과 독일 뮌헨공대 연구진이 개발한 'FaceForensics++' 데이터를 사용하여 연구의 객관성을 높였다. FaceForensics++의 원본 데이터는 현실적인 환경을 위해 유튜브(Youtube)에서 수집하였고, 수집된 영상 속 얼굴이 가려지는 데이터는 수동으로 선별하였다. 최종적으로 509,914개의 이미지를 포함하는 1,000개의 비디오를 선택하여 사용한다.

수집된 1,000개의 데이터를 4가지 딥페이크 기술을 통해 4,000개의 forensic video를 구축하였다. 학습 데이터, 검증 데이터, 테스트 데이터 또한 [10] 논문에서 주어진 지표를 이용하여 필드별 720, 140, 140개로 나누어 총 3,600개의 학습 데이터, 720개의 검증 데이터, 720개의 테스트 데이터로 분할한다.

영상은 평균적으로 1초에 약 30장의 프레임이 존재한다. 학습 시에는 영상별로 10장의 프레임을 추출하여 사용하였고, 테스트 시에는 300장의 프레임을 추출하여 사용하여 영상을 약 10초간 분석 후 딥페이크 조작 여부를 판단

한다.

4.2 FaceDetecor 비교 실험

본 실험에서는 얼굴 탐지 모델에 따른 정확도 변화를 비교를 통해 본 연구에서 사용할 얼굴 탐지 모델을 결정하였다. 정확한 비교를 위해 사용한 모델은 Efficientnet-b4로 통일하였고, 얼굴 탐지를 제외한 정규화 방식, 하이퍼 파라미터 등 학습 조건은 동일한 상태에서 진행하였다. <표 1>은 각각 MTCNN[17], Fast MTCNN[32], BlazeFace[18]로 추출한 얼굴 정보를 이용해 Efficientnet-b4로 Classification을 진행한 결과이다. Efficientnet-b4[27]는 Imagenet 데이터셋[33]으로 사전 학습되어있기 때문에, 추출한 얼굴값은 Imagenet의 평균과 표준편차인 평균 [0.485, 0.456, 0.406], 표준편차 [0.229, 0.224, 0.225]를 사용하여 정규화를 진행하였다.

얼굴 탐지 모델에 따른 정확도는 <표 1>과 같다. 세 가지 방식 모두 평균 정확도에서는 비슷한 성능을 보였으나, 임계 값에 따른 분류 모델의 성능을 보여주는 ROC(Receiver Operating Characteristic) 곡선의 아래 영역을 나타내는 AUC (Area Under the Curve)에서는 성능 차이를 보였다.

BlazeFace는 딥페이크 조작이 이루어진 데이터에 대해서는 높은 성능을 보였으나, 유튜브 데이터 즉, 원본 데이터에 대해서는 정확도가 낮은 수치를

보여 얼굴을 제대로 탐지해내지 못했다는 것을 알 수 있었다. 이에 반해 MTCNN과 Fast MTCNN은 원본 데이터와 조작된 데이터 간 정확도 차이가 크지 않았다. 데이터 불균형으로 인해 원본 데이터의 수가 현저하게 적음에도 불구하고, 유튜브 데이터에 대해 높은 정확도를 보이고, 높은 AUC를 달성한 MTCNN을 본 연구의 얼굴 추출 모델로 사용하고자 한다.

<표 1> Face Detecor 실험 결과표

Dataset	MTCNN	Fast MTCNN	BlazeFace
Youtube	0.91	0.79	0.25
DeepFake	0.86	0.99	0.96
Face2Face	0.74	0.85	0.91
FaceSwap	0.71	0.83	0.95
NeuralTexture	0.58	0.66	0.83
Average Acc	0.76	0.82	0.78
AUC	**0.93**	**0.91**	**0.66**

4.3 정규화 비교

본 실험에서는 이미지 데이터를 전처리하는 방식에 따른 정확도 비교를 통해 본 연구에서 사용할 데이터 전처리 방식을 결정하였다. 정확한 비교를 통해 Efficientnet-b4에 Bi-LSTM Attention을 추가한 모델을 동일하게 사용하였으며, 정규화 방식을 제외한 다른 학습 조건은 동일한 상태에서 진행하였다.

실험을 진행한 정규화 방식은 다음과 같다. 첫 번째, 이미지 데이터를 -1부터 1까지의 범위로 변환시켰다. 일반적으로 사용되는 평균 [0.5, 0.5, 0.5], 표준편차 [0.5, 0.5, 0.5]를 이용해 즉, z-score 정규화를 진행하였다. 두 번째, 이미지 데이터를 0부터 1까지의 범위로 변환시켰다. Imagenet 데이터의 평균과 표준편차인 평균 [0.485, 0.456, 0.406], 표준편차 [0.229, 0.224, 0.225]를 통해 정규화를 진행했다.

두 가지 실험 조건 모두 Threshold를 조정하지 않고 0.5를 기준으로 정상과 조작을 분류한 정확도는 <표 2>와 같다. 모든 데이터, 평균 정확도, AUC 부분에서 모두 [0, 1]'로 정규화 한 수치가 높았다. 본 실험을 통해 본 연구에서는 평균 [0.485, 0.456, 0.406], 표준편차 [0.229, 0.224, 0.225]를 통해 데이터를 0과 1 사이의 값을 갖도록 하는 정규화 방식을 선택했다.

<표 2> 정규화 방식 비교 실험 결과표

	[-1,1] 정규화	[0,1] 정규화
Youtube	0.76	0.78
DeepFake	0.94	0.99
Face2Face	0.93	0.98
FaceSwap	0.96	0.96
Neural Texture	0.87	0.96
Average Acc	0.89	0.934
AUC	0.94	0.954

4.4 제안 모델 비교 실험

본 실험에서는 Single Stream의 정보만 활용한 경우와 Two - Stream 정보를 활용한 경우를 비교함으로써 제안한 방법론의 타당성을 입증한다. 또한,

프레임 간 시간적 정보를 추출하여 불일치성을 파악하는 Bi-LSTM에 Attention 유무를 실험함으로써 시간적 정보에 유익한 정보를 선택적으로 강조하는 Attention의 필요성을 입증한다. 데이터에 대한 전처리는 4.2절과 동일하게 진행했고, 학습에 활용되는 최적화 알고리즘은 Adam을 이용했으며[31], 모델의 출력값과 딥페이크 조작 여부 간 계산되는 손실 함수는 Binary Cross-entropy 값으로 설정하였다.

$$\mathrm{BCE(x)} = -\frac{1}{N}\Sigma_{i=1}^{N} y_i \log\big(h(x_i;\theta)\big) + (1-y_i)\log\big(1-h(x_i;\theta)\big) \quad (3)$$

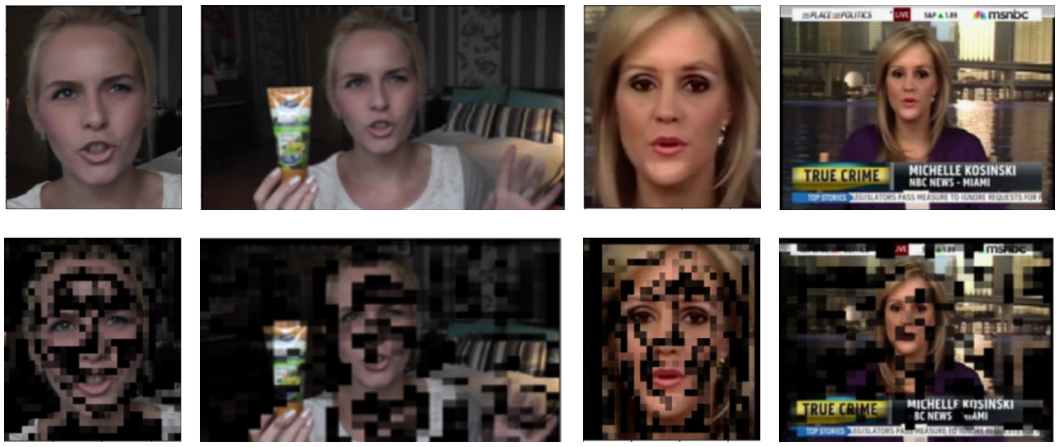
실험 결과는 <표3> 와 같다. Efficient-b4 + Bi-LSTM 모델을 제외하고 모든 모델에 대하여 전역 정보나 지역 정보를 단독으로 사용할 때보다는 두 가지 정보를 모두 활용할 때의 AUC 값이 높음을 통해 더 좋은 성능을 나타냄을 확인할 수 있다. 그뿐만 아니라 단순히 두 가지 정보를 모두 활용하는 Two-stream 모델의 성능보다는 Attention을 추가하였을 때가 사용하지 않았을 때보다 성능이 높은 것을 확인할 수 있다. 또한 EfficientNet[27]의 실험 결과를 기준으로 파라미터 수 대비 가장 좋은 성능을 보이는 Efficientnet-b4와 Efficient-b5 모델로 비교한 결과 전역 정보만 사용할 때의 성능은 b4의 AUC가 높지만, 그 외에는 b5모델의 AUC 값이 높은 것을 통해 더 좋은 정확도를

보이는 결과로 확인 할 수 있다.

<표 3> 제안 모델 비교 실험 AUC 성능 결과

	Single-Stream (Global Information)	Single-Stream (Local Information)	Two-Stream
Efficientnet-b4+Bi-LSTM	**0.901**	0.947	0.944
Efficientnet-b4+ Bi-LSTM +Attention	0.886	0.889	0.954
Efficientnet-b5+Bi-LSTM	0.838	0.953	0.983
Efficientnet-b5+ Bi-LSTM +Attention	0.846	**0.949**	**0.957**

전역 정보와 지역 정보 각각에 사용한 Attention의 영향력을 분석하기 위해 <그림 4>와 같이 시각화하였다. 주어진 정보는 윗줄에 위치하고, 아랫줄은 Attention을 적용한 결과이다. 지역 정보에서 Attention은 조작 발생 시 불일치성이 발생하는 눈, 코, 입에 집중되어 정보를 반영하는 것을 확인했다. 전역 정보에서는 반면에 얼굴과 주변 정보를 균형 있게 모두 활용하여 딥페이크 탐지에 필요한 특징을 추출했음을 확인할 수 있다.



<그림 4> 전역 정보와 지역 정보에 Attention 적용 결과

결과적으로 전역 정보와 지역 정보를 따로 사용했을 때 보다 두 가지 정보를 함께 활용하여 조작된 비디오의 특징을 파악하고, 시점 별로 추출된 특징에 가중치를 부여하여 모든 정보를 이용하기보다는 필요한 정보만을 선택적으로 활용하여 딥페이크 탐지를 효과적으로 수행함을 알 수 있다. 이를 통해 이 연구에서 제안하고자 하는 Deepfake detection model by two stream bi-LSTM attention 방법론의 효과성을 입증하였고, Deepfake detection에 있어 강건하고 확장된 모델이라 할 수 있다.

5. 결론

본 논문에서는 딥페이크 기술로 생성된 영상을 탐지하기 위해 합성곱 신경망(Convolutional neural network ,CNNs) 와 순환 신경망(Recurrent neural networks , RNNs)을 융합하여 사용하였다. 그 과정에서 효과적으로 특징 벡터를 추출하기 위해 전역 정보와 지역 정보, 두 가지 정보를 활용한 딥페이크 탐지 모델을 제안했다. 전역 정보는 동영상에서 추출한 프레임 전체 정보를 사용하고, 지역 정보는 프레임 속 얼굴을 추출한 얼굴 정보만을 사용한다. 얼굴 추출을 위해 다양한 얼굴 추출 모델들을 비교 후 MTCNN을 사용하여 얼굴을 추출했고, 시각적 특징 벡터는 Efficientnet-b4, 시간적 특징은 Bi-LSTM을 활용하여 추출했다. 각 과정에서 딥페이크 탐지에 유용한 특징은 강조하고, 덜 유용한 특징은 억제하기 위해 각 과정에 Attention-Mechanism을 추가했다. 각각의 특징에 따른 실험을 통하여 제안된 모델의 효과성과 우수성을 입증하였다. 본 연구를 통해 효과적으로 딥페이크 기술로 조작된 영상을 탐지해내어, 향후 영상 조작으로 인해 발생하는 문제들을 사전에 방지할 수 있을 것으로 기대된다.

<참 고 문 헌>

[1] R. Chawla, "Deepfakes : How a pervert shook the world." international journal of advance reserch and development, vol 4, issue 6, p.4-8, 2019.

[2] Tormod dag Fikse, "Imaging Deceptive Deepfakes An ethnographic exploration of fake videos" Master's thesis University of Oslo, Oct, 2018.

[3] FINANCIAL TIMES(2019.10.10.)“Deepfakes: Hollywood’s quest to create the perfect digital human”

[4] Y. Li, M. Chang, and S. Lyu, “In ictu oculi : Exposing ai created fake videos by detecting eye blinking.” IEEE Workshop on Information Forensics and Security, Hong Kong, Dec. 2018

[5] D. Wen, H. Han, and A. K. Jain, “Face spoof detection with image distortion analysis.” IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp.

[6] D. Guera and E. J. Delp, “Deepfake video detection using recurrent neural networks.” Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, p. 1-6, Nov. 2018

[7] Hasam Khalid, Simon S. Woo, “OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 656-657

[8] DM Montserrat, H Hao, SK Yarlagadda, “Deepfakes Detection with Automatic Face Weighting”, 2020 arxiv.org

[9] Hua Qi, “DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms”, MM '20: Proceedings of the 28th ACM International Conference on MultimediaOctober 2020 Pages 4318-4327

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” Proceedings of the IEEE International Conference on Computer Vision, pp. 1-11, October 2019, Seoul, South Korea.

[11] “DeepFakes,” https://github.com/deepfakes/ faceswap. 2

[12] J. Thies, M. Zollhofer, and M. Nießner, “Deferred neural ¨ rendering: Image synthesis using neural textures,” ACM Transactions on Graphics, vol. 38, no. 4, pp. 1-12, July 2019.

[13] M. Kowalski, “Faceswap,” https://github.com/ MarekKowalski/FaceSwap/.

[14] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, June 2016, pp. 2387-2395.

[15] KARRAS, Tero; LAINE, Samuli; AILA, Timo. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2019. p. 4401-4410.

[16] Shan Li, Weihong Deng “Deep Facial Expression Recognition : A

Survey" 2018 1st SIBGRAPI Conference on Graphics, Patterns and Images, pp.471-478, Oct.2018.

[17] ZHANG, Kaipeng, et al. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016, 23.10: 1499-1503.

[18] BAZAREVSKY, Valentin, et al. Blazeface: Sub-millisecond neural face detection on mobile gpus. arXiv preprint arXiv:1907.05047, 2019.

[19] KORSHUNOV, Pavel; MARCEL, Sébastien. Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685, 2018.f

[20] GALBALLY, Javier; MARCEL, Sébastien. Face anti-spoofing based on general image quality assessment. In: 2014 22nd international conference on pattern recognition. IEEE, 2014. p. 1173-1178.

[21] GÜERA, David; DELP, Edward J. Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018. p. 1-6.

[22] NGUYEN, Huy H.; YAMAGISHI, Junichi; ECHIZEN, Isao. Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467, 2019.

[23] MONTSERRAT, Daniel Mas, et al. Deepfakes Detection with Automatic Face Weighting. arXiv preprint arXiv:2004.12027, 2020.

[24] WU, Xi, et al. SSTNet: Detecting Manipulated Faces Through Spatial, Steganalysis and Temporal Features. In: ICASSP 2020-2020 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020. p. 2952-2956.

[25] “Deepfake Detection Challenge (DFDC),” https://deepfakedetectionchallenge.ai/, 2019.

[26] “FaceForensics Benchmark,” http://kaldir.vc.in.tum.de/faceforensics_benchmark/, 2019.

[27] TAN, Mingxing; LE, Quoc V. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946, 2019.

[28] PARK, Jongchan, et al. Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514, 2018.

[29] SCHUSTER, Mike; PALIWAL, Kuldip K. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 1997, 45.11: 2673-2681.

[30] BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[31] KINGMA, Diederik P.; BA, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[32] “Fast-MTCNN “, https://github.com/szad670401/Fast-MTCNN

[33] Deng, J. et al., 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. pp. 248-255.

[34] BBC News, ‘The Fake Video Where Johnson and Corbyn Endorse Each

Other' (BBC News, 12 November 2019) <https://www.bbc.co.uk/news/av/technology-50381728/the-fake-video-where-johnson-and-corbyn-endorse-each-other>.

[35] P. Korshunov and S. Marcel, "Speaker inconsistency detection in tampered video." European Signal Processing Conference(EUSIPCO), p. 2375-2379, Sep. 2018.

[36] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[37] BONETTINI, Nicolò, et al. Video Face Manipulation Detection Through Ensemble of CNNs. arXiv preprint arXiv:2004.07676, 2020.

[38] 이승환, 빅데이터로 본 딥페이크(Deepfake) : 가짜와의 전쟁. SPRi-소프트웨어정책연구소, 2020

<영 문 요 약>

ABSTRACT

Two-Stream Attention-Based Efficientnet and Bi-LSTM Networks for DeepFake Detection

Nayeon Song

Dept. of Industrial Engineering

The Graduate School

Yonsei University

Recently, deep learning has been successfully applied and used in various fields. Especially in the field of computer vision, the development of deep learning has produced many conveniences. However, as technology advances, the number of abuse cases of technology threatening personal privacy and national security increases proportionally. One of these technologies is Deepfake. Deepfake is also used in various fields such as movies and records due to its high industrial value. However,

many problems are being raised that exploit distorted facts through video manipulation. As the number of cases of abuse of Deepfake increases, research is actively being conducted to determine whether a video is manipulated or not.

This study aims to accurately detect Deepfake manipulation by extracting visual information in the frame and temporal information between the frames using convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In previous studies, there is a limitation that time information between frames cannot be included using CNNs but only using visual information. Furthermore, recent studies only detect Deepfake images by limiting facial information in frames, so there is a limit to detecting only facial parts where manipulation occurred.

This study proposes more effective Deepfake video detection by extracting features from two sources of information in the video: global and local. Local information includes information on the face area where manipulation occurs, and global information includes the complete frame information that can determine the harmony of the entire video. Therefore, the proposed detection model demonstrates the results of the convergence

of these sources of information to complement the deficiencies. The proposed model is evaluated using the FaceForces++ datasets and is competitive with other models.

Keyword: Manipulated video detection, Deepfake detection, Efficientnet, Video forensics detection, Two-stream;