Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.30, NO.6, Dec. 2020 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2020.30.6.1053

Bidirectional Convolutional LSTM을 이용한 Deepfake 탐지 방법

이 대 현,[1†] 문 종 섭[2‡]
[1,2]고려대학교 (대학원생, 교수)

A Method of Detection of Deepfake Using Bidirectional Convolutional LSTM

Dae-hyeon Lee,[1†] Jong-sub Moon[2‡]
[1,2]Korea University (Graduate student, Professor)

요 약

최근 하드웨어의 성능과 인공지능 기술이 발달함에 따라 육안으로 구분하기 어려운 정교한 가짜 동영상들이 증가하고 있다. 인공지능을 이용한 얼굴 합성 기술을 딥페이크라고 하며 약간의 프로그래밍 능력과 딥러닝 지식만 있다면 누구든지 딥페이크를 이용하여 정교한 가짜 동영상을 제작할 수 있다. 이에 무분별한 가짜 동영상이 크게 증가하였으며 이는 개인 정보 침해, 가짜 뉴스, 사기 등에 문제로 이어질 수 있다. 따라서 사람의 눈으로도 진위를 가릴 수 없는 가짜 동영상을 탐지할 수 있는 방안이 필요하다. 이에 본 논문에서는 Bidirectional Convolutional LSTM과 어텐션 모듈(Attention module)을 적용한 딥페이크 탐지 모델을 제안한다. 본 논문에서 제안하는 모델은 어텐션 모듈과 신경곱 합성망 모델을 같이 사용되어 각 프레임의 특징을 추출하고 기존의 제안되어왔던 시간의 순방향만을 고려하는 LSTM과 달리 시간의 역방향도 고려하여 학습한다. 어텐션 모듈은 합성곱 신경망 모델과 같이 사용되어 각 프레임의 특징 추출에 이용한다. 실험을 통해 본 논문에서 제안하는 모델은 93.5%의 정확도를 갖고 기존 연구의 결과보다 AUC가 최대 50% 가량 높음을 보였다.

ABSTRACT

With the recent development of hardware performance and artificial intelligence technology, sophisticated fake videos that are difficult to distinguish with the human's eye are increasing. Face synthesis technology using artificial intelligence is called Deepfake, and anyone with a little programming skill and deep learning knowledge can produce sophisticated fake videos using Deepfake. A number of indiscriminate fake videos has been increased significantly, which may lead to problems such as privacy violations, fake news and fraud. Therefore, it is necessary to detect fake video clips that cannot be discriminated by a human eyes. Thus, in this paper, we propose a deep-fake detection model applied with Bidirectional Convolution LSTM and Attention Module. Unlike LSTM, which considers only the forward sequential procedure, the model proposed in this paper uses the reverse order procedure. The Attention Module is used with a Convolutional neural network model to use the characteristics of each frame for extraction. Experiments have shown that the model proposed has 93.5% accuracy and AUC is up to 50% higher than the results of pre-existing studies.

Keywords: Deepfake, LSTM, Attention module, Artificial intelligence, Time distribution.

Received(10. 08. 2020), Modified(11. 13. 2020), Accepted(11. 13. 2020)

† 주저자, msms951009@gmail.com
‡ 교신저자, jsmoon@korea.ac.kr(Corresponding author)

I. 서 론

최근 하드웨어의 성능과 인공지능 기술이 크게 향상됨에 따라 얼굴 합성 기술은 사람의 인지력으로는 진위를 가리기 힘들 정도로 발전하였다. 인공지능을 이용한 얼굴 합성 기술을 딥페이크라 하며 딥페이크를 이용해 제작한 가짜 동영상 또한 딥페이크라는 용어로 일반화되었다. 딥페이크는 "딥페이크"라는 아이디를 사용하는 사용자가 텐서플로우(TensorFlow)와 같은 딥러닝 오픈소스 소프트웨어를 이용해 유명 연예인과 포르노를 합성하면서 유명해졌다. 해당 가짜 동영상은 구글 이미지 검색, 유튜브 등 쉽게 획득 가능한 데이터를 활용하여 제작되었으며 이는 약간의 프로그래밍 능력과 딥러닝 지식만 있으면 누구든지 쉽게 가짜 동영상을 만들 수 있음을 보여주었다. 이후 인공지능 기술을 이용한 FakeApp[38]이라는 무료 앱이 출시되면서 무분별한 많은 가짜 동영상이 생성되었다. 2018년 이후부터 정치인들과 배우들의 가짜 뉴스와 포르노 영상들로 인해 딥페이크가 사회적 문제로 부각 되었다.

이와 같이 딥페이크 기술로 인해 사회적으로 많은 문제가 발생하고 있는 반면, 현재까지 딥페이크에 대한 뚜렷한 대책과 대응 방안은 제시되지 않았다. 이에 가짜 동영상으로 인한 피해를 예방하고자 본 논문에서는 딥페이크 영상을 탐지하는 모델을 제안한다.

본 논문에서 제안하는 딥페이크 탐지 모델의 기여는 다음과 같다. 첫째, 본 논문에서 제안하는 모델이 기존 제안된 모델들이 탐지하지 못하던 데이터 셋을 학습하여 높은 성능을 보인다. 둘째, 프레임간 지속적으로 나타나는 불연속적인 위치를 특정하여 후속 연구의 학습범위로 제안한다. 마지막으로 기존 Fully-connected layer와 LSTM 모델을 단순하게 연결한 FC-LSTM 기반 딥페이크 탐지 모델보다 공간적, 시간적 특징을 효율적으로 학습하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 딥페이크 탐지 관련 연구로써 기존에 제안된 딥페이크 탐지 모델의 특징과 연구에 사용된 데이터 셋에 대해 설명한다. 3장에서는 본 논문에서 제안하는 방법에 대해 전체적인 구조와 단계를 자세하게 설명한다. 4장에서는 본 논문에서 사용한 딥페이크 데이터 셋과 제안하는 모델을 이용한 딥페이크 데이터 셋을 분류한 결과의 정확도와 탐지율을 비교 분석하였다. 5장에서는 결론에 관해서 기술하였다.

II. 관련 연구

본 장에서는 기존에 사용되던 여러 가지 얼굴 이미지에 대한 조작방법과 이를 탐지하는 딥페이크 탐지 방법에 대해 설명한다.

2.1 얼굴 이미지 조작기법 (Face Image Manipulation Methods)

이미지 조작기법은 크게 전통적으로 그래픽 기반 얼굴 조작 프로그램을 이용한 방법과 딥러닝을 이용한 방법으로 분류된다.

그래픽 기술 기반 얼굴 조작은 3D 다중 선형 모델을 이용하여 실시간으로 원본 동영상의 얼굴을 다른 사람의 얼굴로 대체하는 방법[28]과 RGB 카메라만을 이용하여 실시간으로 동영상의 얼굴을 조작하는 Face2Face[29]가 있다. 3D 다중 선형 모델을 이용한 방법은 기존 이미지 조작방법들과 다르게 사용을 위한 준비와 특별한 하드웨어가 요구되지 않는 장점이 있지만, 얼굴 위로 조명이 천천히 바뀌어야 하고 고주파 조명이 존재하는 경우 이미지 조작이 정상적으로 이뤄지지 않는 단점이 있었다. Face2Face에 경우, 처음으로 발표된 monocular RGB 입력만으로 실시간으로 얼굴을 조작하여 AR/VR 분야 등에 새로운 기반이 될 수 기술이다. 하지만 긴 머리와 수염이 있는 얼굴과 비교적 짧은 영상에서는 경우 충분한 성능을 내지 못하는 단점이 있다.

딥러닝 기반으로 얼굴을 조작하는 기술은 딥페이크가 대표적이다. 딥페이크는 생산적 적대 신경망(Generative Adversarial Networks, GAN)[36] 또는 오토인코더(Autoencoder)[34, 35]를 기반으로 얼굴을 조작하는 인공지능 기술이다. 딥페이크가 등장하면서 많은 응용 프로그램이 만들어졌으며 ZAO에서 만든 FaceApp[37]은 인공지능 기술을 기반으로 손쉽게 동영상 내 사람의 얼굴을 조작할 수 있는 스마트폰 애플리케이션이다. FaceApp은 사진에서 매우 사실적인 다른 얼굴을 자동으로 생성한다. 스마트폰을 이용해 머리 모양, 성별, 나이 등을 바꿀 수 있지만, 영상이 아닌 사진만 조작할 수 있는 단점이 있다.

2.2 Deepfake 탐지 연구

딥러닝을 이용한 딥페이크 탐지는 크게 동영상의

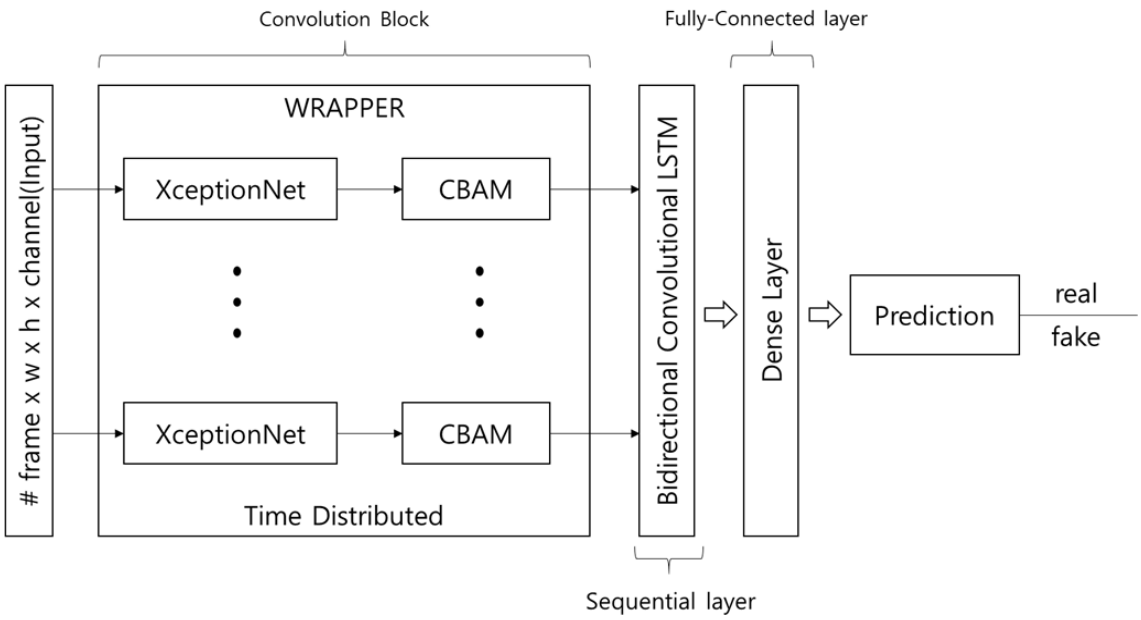


Fig. 1. Overall structure of proposed method

프레임별로 특징을 학습하여 가짜 동영상을 탐지하는 "Visual Artifacts within Frame"과 인접한 프레임에 대해서 연결이 부자연스러운 특징을 학습하는 "Temporal Features across Frames"로 나뉜다.

"Visual Artifacts within Frame"은 동영상을 프레임 단위로 나누고 합성곱 신경망(Convolutional Neural Network, CNN)을 통해 실제 동영상과 가짜 동영상의 각 프레임의 특징을 학습한다. 합성곱 신경망은 이미지 분석에 주로 적용되는 인공신경망 모델 중 하나이다. 가짜 동영상의 경우 각 프레임의 전체적인 자연스러움을 고려하지 않고 프레임마다 다른 사람의 얼굴로 바꾸기 때문에 얼굴과 얼굴 주변의 배경 사이에 부자연스러움이 발생한다. "Visual Artifacts within Frame"은 가짜 동영상에서 발생할 수 밖에 없는 부자연스러움을 합성곱 신경망을 통해 학습한다. 실제 동영상과 가짜 동영상의 각 프레임을 합성곱 신경망을 통해 학습하여 추출한 각 프레임의 특징을 바탕으로 입력된 프레임이 실제 동영상인지 가짜 동영상인지를 판별한다

합성곱 신경망 모델인 XceptionNet[30], InceptionV3[31], SeNet[32], ResNet[33]은 이미지의 특징 추출에 가장 많이 사용되는 기법이다. 최신 딥페이크 탐지 연구에 따르면 합성곱 신경망만을 사용하지 않고 어텐션 모듈(Attention module)을 같이 사용하는 것이 더 높은 정확도를 보였다[21]. 어텐션 모듈을 합성곱 신경망과 같이 사용할 경우, 학습에 필요한 부분에 집중하여 정확도를 향상시킬 수 있지만, 어텐션 모듈이 비지도 학습이기 때문에 충분히 복잡한 신경망 모델이 아니면 어텐션 모듈을 사용할 경우 더 비효율적일 수 있다.

"Temporal Features across Frame"은 순환 신경망(Recurrent Neural Network, RNN)을 통해 각 프레임의 연결에 초점을 맞추어 학습한다. 순환 신경망은 시계열 데이터와 같이 시간의 흐름에 따라 변화하는 데이터를 학습하기 위한 인공신경망이다. 순환 신경망을 통해 인접한 프레임에 대해 프레임 간 연결이 자연스럽지 않고 일관적이지 않은 특징을 찾아서 실제 동영상인지 가짜 동영상인지를 판별한다.

최근 딥페이크 탐지 연구는 순환 신경망을 개선한 Long Short-Term Memory(LSTM) 셀을 이용하는 추세이다[27]. LSTM 셀은 데이터의 시간과 순서를 고려할 수 있어 프레임간 불연속적인 특징을 학습 시 적합하다. 하지만 LSTM 셀 자체에 단점을 그대로 적용된다는 문제점이 있다. LSTM 셀은 학습하는데 많은 시간이 걸리고, 많은 하드웨어 자원(메모리), 많은 양의 데이터가 필요한 단점이 있다.

III. 제안 방법

3.1 Deepfake 탐지 모델 개요

본 논문에서는 양방향 순환 신경망을 활용한 딥페이크 탐지 모델을 제안한다.

Fig. 1.과 같이 본 논문에서 제안하는 모델은 Convolution Block, Sequential layer, Fully-Connected layer로 구성된다.

본 논문에서 제안하는 모델은 딥페이크 데이터 셋을 학습한다. 전 처리된 데이터 셋을 이용하여 Convolution Block에 입력한다. Convolution Block 내 n개의 출력이 Sequential layer를 구성하는 각 셀로 순차적으로 입력된다. Sequential layer의 최종 출력이 Fully-connected layer의 입력이 된다. Fully-connected layer는 Softmax 함수를 사용하여 출력 노드 두 개로 학습한다. 결과적으로 실제 동영상인지 딥페이크 영상인지를 카테고리 판정한다.

Convolution Block에서는 프레임 단위로 분할된 동영상의 이미지 데이터로부터 각 프레임의 특징을 추출한다. Sequential layer는 프레임 간 연결 관계를 학습한다. Fully-Connected layer에는 입력된 동영상이 딥페이크로 생성된 가짜 동영상인지 실제 동영상인지를 판별한다.

3.2 Convolution Block

본 논문에서는 동영상의 프레임별 특징을 추출하기 위해 합성곱 신경망 모델 중 하나인 XceptionNet과 중요한 특징에 가중치를 주어 학습하는 어텐션 모듈을 이용한다.

우선 프레임 간의 연결 관계를 학습시키기 위해 XceptionNet을 사용하여 각 프레임의 특징을 추출한다. XceptionNet은 Inception V3보다 파라미터 개수를 줄여 효율성과 성능을 높인 모델이다[30] XceptionNet은 다른 합성곱 신경망 모델과 비교하였을 때, 비슷한 파라미터 수를 가지는 모델 중에서 좋은 성능 가지고 있고 효율적으로 이미지의 특징을 추출할 수 있다[30, 41, 42] 특히 딥페이크 탐지 모델 중에서 XceptionNet을 이용한 모델들이 가장 좋은 성능을 가진다[26]

이후 XceptionNet으로 추출된 프레임의 특징을 Convolutional Block Attention Module(CBAM)[12]에 전달하여 CBAM을 학습시킨다.

CBAM은 어텐션 모듈 중의 하나로써 가짜 동영상 판별에 큰 영향을 미치는 특징에 가중치를 주어 모델의 정확도를 높인다.

CBAM은 Channel attention, Spatial attention을 차례대로 적용한다. Channel attention module에서는 이미지 내에 "무엇"이 의미 있는 특징인가에 집중하며, Spatial attention module에서는 이미지 내 "어디"가 유익한 부분인지에 집중한다. 학습된 CBAM을 통해 본 논문에서 제안한 모델은 입력된 동영상이 가짜 동영상인지 실제 동영상인지 판별할 때 영향을 준 이미지의 영역을 강조할 수 있다[12] CBAM에서 말하는 "무엇"이란 딥러닝 모델이 판별한 물체를 의미하여, 이미지 내의 "어디"란 판별한 물체의 위치를 의미한다. 예를 들어 Fig. 2.에서 "무엇"이란 사람을 의미하여 "어디"란 사람의 위치를 의미한다.

XceptionNet으로부터 추출된 특징의 크기는 식(1)으로 표시할 수 있다. 식(1)에서 "H"는 높이, "W"는 폭, "C"은 채널 수를 의미한다. 식 (2)~(4)는 XceptionNet으로부터 추출된 특징에 CBAM을 적용하는 과정을 표현한 것이다.

$$F \in \mathbb{R}^{H \times W \times C} \tag{1}$$

$$F' = M_c(F) \otimes F \tag{2}$$

$$F'' = M_s(F') \otimes F' \tag{3}$$

$$\overline{F} = F \oplus F'' \tag{4}$$

Fig. 3.와 같이 XceptionNet으로부터 추출된 특징 F를 CBAM의 구성요소인 Channel attenti

Fig. 2. Example of Image

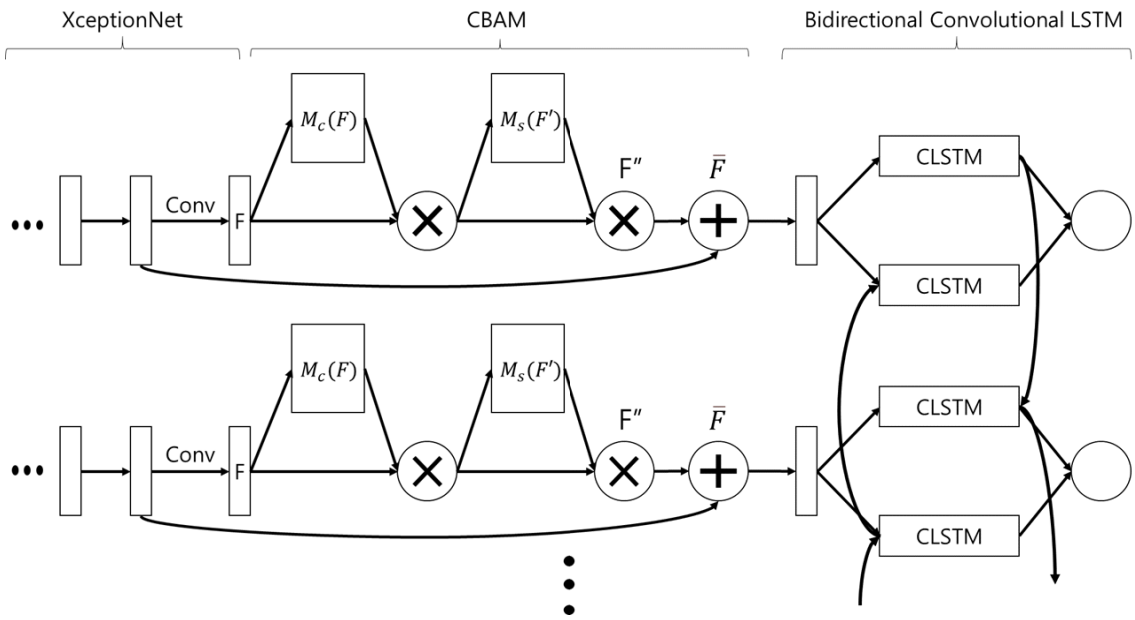


Fig. 3. Detailed structure of the proposed model

on module "M_c"과 Spatial attention module "M_s"에 순서대로 적용한다. "$\oplus$"은 element-wise 합, "$\otimes$"은 아다마르 곱 연산이다.

XceptionNet으로부터 추출된 특징을 CBAM의 첫 번째 구성요소인 M_c에 적용하여 모델을 학습시키는 과정은 식 (5)과 같이 계산한다.

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) \\ &\quad + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F_{avg}^C)) + W_1(W_0(F_{\max}^C))) \end{aligned} \tag{5}$$

식 (5)에서 "σ"은 시그모이드 함수를 의미하며, "W_0", "W_1"은 다층 퍼셉트론(Multi-Layer Perceptron, MLP)의 가중치이다.($W_0 \in \mathbb{R}^{C/r \times C}$, $W_1 \in \mathbb{R}^{C \times C/r}$, r은 오버헤드를 줄이기 위한 축소율)

$$\begin{aligned} M_s(F') &= \sigma(f^{N\times N}([AvgPool(F'); \\ &\quad MaxPool(F')])) \\ &= \sigma(f^{N\times N}(F'^s_{avg}; F'^s_{avg})) \end{aligned} \tag{6}$$

식 (6)은 CBAM의 두 번째 구성요소인 Spatial attention module "M_s"을 적용하는 계산식이다. 식 (6)에서 "σ"은 시그모이드 함수를 의미하며, "$f^{N\times N}$"은 필터 크기가 $N\times N$인 컨볼루션 연산자이다.

각 XceptionNet-CBAM은 프레임간 사이의 관계를 학습하기 위한 Sequential layer와 연결되기 위해서 Time distributed layer 형태로 감싸져 있다. $m \times F \times W \times H \times C$("$m$"은 배치 크기) 형태로 입력 데이터가 본 논문에서 제안하는 모델에 입력되어 각 프레임별로 XceptionNet-CBAM을 통해 특징이 추출된다. 그 후, 모든 프레임의 특징들이 Sequential layer에 공급된다. "F"은 프레임의 개수로 "F"의 크기만큼 Xception-CBAM의 개수를 설정하고, "W", "H"은 프레임의 폭, 높이로 학습하는 프레임의 크기에 따라 결정된다. "C"는 채널의 개수이다.

Fig. 3.와 같이 CBAM에서 나오는 각각의 출력은 연결된 두 개의 Convolutional LSTM(CLSTM) 셀에 연속적으로 입력된다. 한 개의 CLSTM 셀은 시간의 순방향을 고려하고, 다른 한 개의 CLSTM 셀은 시간의 역방향을 고려한다. 두 개의 CLSTM 셀은 하나의 출력이 나오는데, 각각의 출력들은 Fully-connected layer를 구성하는 노드가 된다.

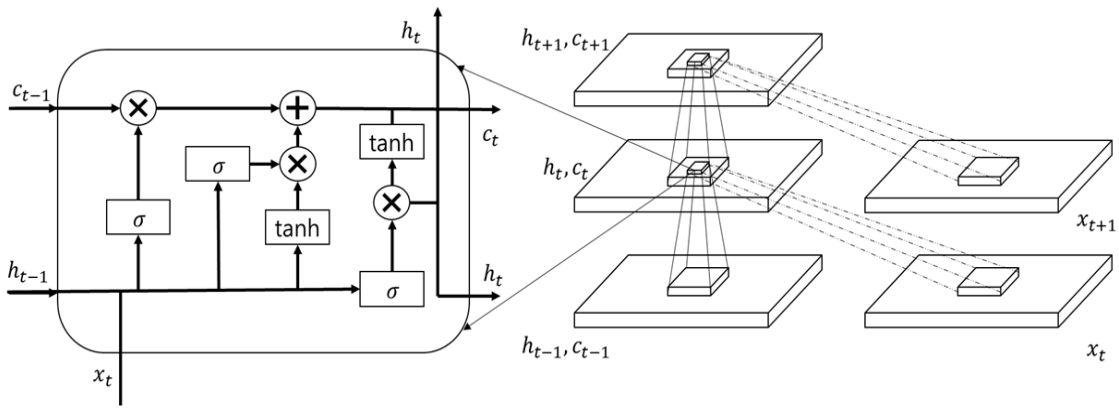


Fig. 4. Structure of the Convolutional LSTM

3.3 Sequential layer

본 논문에서 제안하는 모델의 구성요소 중 Sequential layer는 Bidirectional Convolutional LSTM을 사용하는데 그 전에 먼저 3.3.1절에서 Bidirectional Convolutional LSTM의 구성요소인 Convolutional LSTM을 먼저 설명한다. Bidirectional Convolutional LSTM은 3.3.2절에서 설명한다.

3.3.1 Convolutional LSTM (CLSTM)

본 논문에서는 인접한 프레임에 대해서 프레임 간 연결이 자연스럽지 않고 일관적이지 않은 특징을 학습하기 위해 Convolutional LSTM(CLSTM) 셀을 사용한 순환 신경망 모델을 활용한다. 순환 신경망 모델은 XceptionNet과 CBAM이 결합한 Convolution Block으로부터 추출된 특징을 입력으로 받아 학습한다.

기존 LSTM은 이미지와 같은 3차원(높이, 폭, 채널) 텐서로는 학습할 수 없다. LSTM은 인코딩되어 있거나 특징이 사전에 추출된 1차원 벡터로만 학습할 수 있다. 따라서 LSTM 셀은 데이터를 인코딩하거나 차원을 줄여 특징을 추출하는 단계를 거쳐야만 학습할 수 있다. 반면, CLSTM 셀은 Fig. 4.과 같이 LSTM 셀에 컨볼루션 연산을 포함시켜 3차원 텐서로 학습할 수 있으므로 부가적인 단계를 걸칠 필요가 없이 공간적, 시간적 특징을 동시에 학습할 수 있다[14]. Fig. 4.에서 표현된 것처럼 Convolutional LSTM은 시간의 흐름에 따라 컨볼루션 연산을 수행한다. 이로써 컨볼루션 연산을 통해 공간적 특징을, LSTM을 통해 시간적 특징을 학습할 수 있다[14]

또한, CLSTM 셀은 Fully-connected layer (FC)와 LSTM 셀을 단순하게 연결한 FC-LSTM 모델보다 학습에 이용하는 가중치 수가 적기 때문에 FC-LSTM보다 효율적이다.

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \otimes c_{t-1} + b_i) \quad (7)$$

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \otimes c_{t-1} + b_f) \quad (8)$$

$$c_t = f_t \otimes c_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \quad (9)$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \otimes c_t + b_o) \quad (10)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (11)$$

식 (7)~(11)은 CLSTM의 식이다. CLSTM 셀은 LSTM 셀과 다르게 컨볼루션 연산을 하는 것이 특징이다. "$\otimes$"은 아다마르 곱 연산이고, "*"은 컨볼루션 연산이다.

Convolution Block을 통해 프레임마다 추출된 feature map을 x_t로 두고 CLSTM에 입력하여 은닉 상태(h_t), 셀 상태(c_t)을 계산한다. 이 과정을 통해 필요 없는 과거 정보를 잊고(식(8)의 f_t), 필요한 현재 정보를 기억하면서(식(9)의

$i_t \circ \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c))$ 연속된 프레임 사이의 특징을 학습한다.

3.3.2 Bidirectional Convolutional LSTM (Bi-CLSTM)

기존 순환 신경망은 시간의 순방향으로만 학습이 가능하였다. 역방향의 연결은 존재하지 않기 때문에 현시점 t보다 미래 시점의 데이터를 학습에 이용하지 못한다. 본 논문에서는 딥페이크 동영상 탐지의 정확도를 높이기 위해 역방향의 연속적인 프레임도 고려한다.

본 논문에서 제안하는 모델은 인접한 프레임의 불연속적인 특징을 학습할 때, 시간의 순방향과 역방향을 학습하기 위해서 Bidirectional Convolutional LSTM(Bi-LSTM) 모델을 이용한다.

Bi-LSTM은 Fig. 5.과 같이 순서대로 데이터를 학습하는 CLSTM 셀과 역 순서로 데이터를 학습하는 CLSTM 셀로 구성된다.

시간 간격이 1부터 t까지 있다고 가정할 때, 순방향 CLSTM 셀은 1부터 t까지 순서대로 Convolution Block으로부터 추출된 feature map을 학습한다. 역방향 CLSTM 셀은 t부터 1까지 거꾸로 Convolution Block으로부터 추출된 feature map을 학습한다.

Fig. 6.와 같이 각 프레임을 Convolution Block의 입력하고, Convolution Block의 각각의 출력

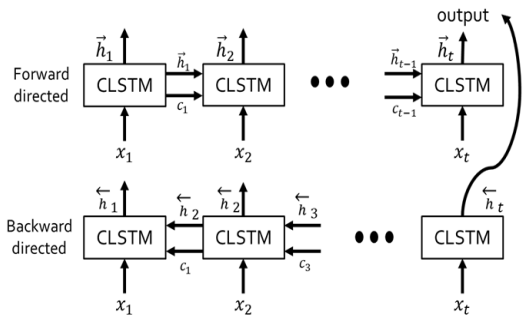


Fig. 5. Structure of the Bi-LSTM

이 시간의 순서대로 두 개의 CLSTM 셀에 입력된다. 두 개의 CLSTM 셀은 시간의 순방향과 역방향을 고려하여 학습한다. CLSTM의 출력은 Fully-connected layer의 노드로 구성된다.

3.4 Fully-connected layer

Fig. 7.와 같이 Fully-connected layer를 추가하여 Sequential layer의 출력을 입력받아 Softmax 함수를 사용하여 입력받아 계산된 값 중 큰 값을 이용하여 실제 또는 가짜 동영상인지 판정한다.

Fully-Connected layer는 퍼셉트론으로 이루어진 층 여러 개를 순차적으로 연결한 형태이다. 가장 기본적인 형태의 인공신경망(Artificial Neural Network) 구조이며, 하나의 입력층(input laye

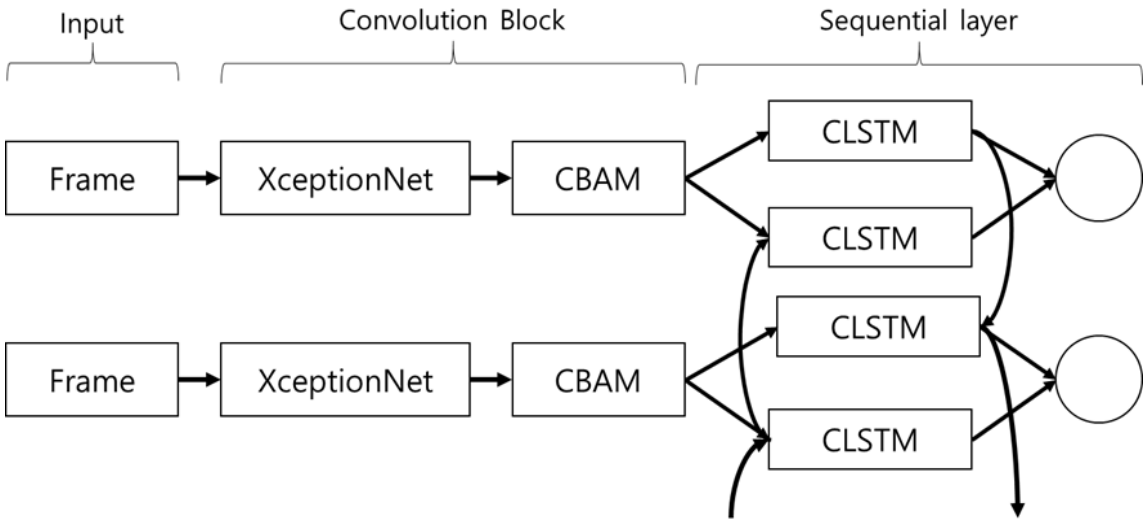


Fig. 6. Detailed structure of the Convolution Block and Sequential layer

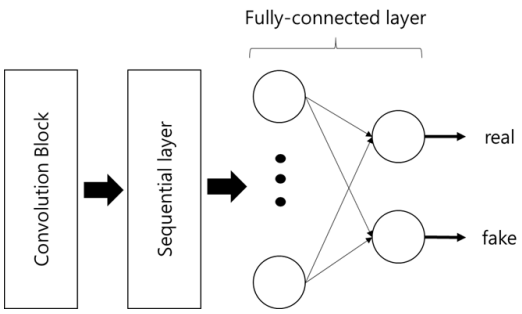


Fig. 7. Structure of the Fully-connected layer

r), 하나 이상의 은닉층(hidden layer), 그리고 하나의 출력층(output layer)로 구성된다.

본 논문에서는 Fully-Connected layer를 분류층으로써 사용한다. Convolution Block, Sequential layer를 거쳐 출력된 특징을 학습하여 딥페이크 영상인지 실제 영상인지 판별한다.

Fully-Connected layer에서는 딥페이크 영상을 탐지하기 위해 두 개의 출력 노드의 활성화 함수로 로지스틱 회귀 모델에서 이진 분류에 사용되는 식(12)와 같은 시그모이드(sigmoid) 함수를 사용한다. 결과값 중 큰 값을 선정하여 실제 동영상인지 가짜 동영상인지를 판정한다.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{12}$$

IV. 실험 결과

이 장에서는 본 논문에서 제안한 딥페이크 탐지 모델의 성능을 측정하기 위해 정확도를 측정하고 그 결과를 제시한다.

4.1 실험 데이터

현재까지 딥페이크 동영상 탐지 연구를 위해 사용되어왔던 데이터 셋은 1세대와 2세대로 분류한다[26]. 1세대 데이터 셋의 구성은 Table. 1.과 같다. 1세대 데이터 셋의 경우 육안으로도 실제 영상과 딥페이크 영상을 판별 가능하였기 때문에 실질적으로 딥페이크 탐지 모델이 효과가 없다. 그러므로 1세대 데이터 셋을 이용한 탐지 모델은 본 논문에서 제안하는 모델과 비교 대상이 되지 않는다.

2세대 데이터 셋은 1세대 데이터 셋의 해상도, 색의 불균형, 부적절한 얼굴 마스킹 작업으로 인한 일그러짐 등의 문제를 개선하였다. 2세대 데이터 셋의 구성은 Table. 2.과 같다.

실험에 사용한 데이터 셋은 딥페이크 탐지에 많이 사용되는 2세대 데이터 셋인 Celeb-DF를 이용하였다. 본 논문에서 제안하는 모델의 기본 구조는 XceptionNet이기 때문에 기존 XceptionNet을 이용한 탐지 모델들과 본 논문에서 제안하는 모델과 비교를 하기 위해서 Celeb-DF 데이터 셋을 이용하였다. 실험에 사용한 데이터 셋의 구성은 Table 3.과 같다. Celeb-DF 데이터 셋의 각 동영상은 30개의 프레임으로 구성되어 있으며 딥페이크 동영상은 유튜브에 업로드된 동영상을 이용하여 생성하였다. 실제 동영상은 총 59명의 인터뷰 영상이며, 성별 및 나이를 균형 있게 선별되어있다. 실험에 사용된 동영상은 인물이 바라보는 방향, 조명, 배경 등이 다양하여 본 논문이 제안하는 딥페이크 탐지 모델에 적합하다.

Table 1. First generation Deepfake data set

Dataset	# Real	# Fake	Release By
UADFV(2018)	49	49	Li et al.
FF-DF(2019)	1,000	1,000	Rossler et al.

Table 2. Second generation Deepfake data set

Dataset	# Real	# Fake	Release By
DFDC(2019)	1,131	4,113	Facebook et al
Celeb-DF(2019)	590	5,639	Li et al.

Table 3. Dataset for Deepfake detection

Set	# Real	# Fake
Training	490	5,539
Test	100	100
Total	590	5,639

4.2 실험 구성

실험은 Celeb-DF 데이터 셋을 학습하여 딥페이크 동영상 탐지를 수행하는 방식으로 진행하였다. Grad-CAM을 이용하여 본 논문에서 제안한 모델이 딥페이크 동영상 판별할 때 영향을 준 구역을 확인하였다.

학습을 위해 6,229개 동영상을 사용하였다. 우선 Celeb-DF 데이터 셋의 각 동영상을 30개의 프레임

으로 분리한 뒤 얼굴만 추출하는 전처리 과정을 진행하였다.

가공된 데이터는 $m \times F \times W \times H \times C$(m은 배치 크기, F는 프레임의 개수, W는 폭, H는 높이, C는 채널 수) 형태로 본 논문에서 제안하는 모델의 입력으로 사용하였다. 실험 시 $F=30$, $W=240$, $H=240$, $C=3$으로 데이터를 구성하였다. CBAM은 필터 크기가 7×7인 컨볼루션 연산자를 이용하고 Table. 4.와 같은 알고리즘으로 학습을 진행하였다. 학습 시, 러닝 레이트(learning rate)는 10^{-4}, 배치 크기는 2로 설정한다. 옵티마이저(optimizer)는 Adam optimizer($\beta_1 = 0.9$, $\beta_2 = 0.999$)을 사용하였다.

본 논문에서 제안하는 모델이 훈련 데이터 셋을 얼마나 잘 처리하지 못하였는지 계산하기 위해 식 (13)과 같이 교차 엔트로피 오차(Cross Entropy Error, CEE)를 이용하였다. 교차 엔트로피 오차 중 이진 분류 문제를 해결하는데 사용하는 binary cross entropy를 사용하였다.

$$E = -\sum_{k} t_k \log y_k \tag{13}$$

Table 4. Learning algorithm of proposed model

Algorithm
inputs
- N is number of real video
- M is number of fake video
- $N < M$
- real video data set $D^r = \{d_1^r, \ldots, d_N^r\}$
- fake video data set $D^f = \{d_1^f, \ldots, d_M^f\}$
outputs
- trained deep neural net weight
1: $c = 0$
2: **while** $c < \frac{M}{N}$ **do**
3: **for** $i \in 1 \ldots N$ **do**
4: $W \leftarrow loadWeight()$
5: $T \leftarrow \{d_i^r, d_{N \times c+i}^f\}$
6: $W \leftarrow LarningModel(W, T)$
7: $SaveWeight(W)$
8: **end for**
9: $c+1$
10: **end while**
11: **return** W

본 논문에서 제안하는 모델의 정확도를 측정하기 위해 식 (14)와 같이 정확도를 측정한다. 식 (14)에서 TP는 True Positive로 실제랑 예측값이 True인 경우, TN은 True Negative로 실제값과 예측값이 False인 경우, FP는 False Positive로 실제값은 False이지만 예측값이 True인 경우, FN은 실제값이 True이지만 예측값이 False인 경우를 의미한다.

$$(Accuracy) = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{14}$$

딥페이크 탐지 모델의 성능을 더 정확하고 자세하게 평가하기 위해서 인식 및 탐지 모델의 성능을 평가하는데 많이 쓰이는 Precision, Recall, F1-score를 계산한다. 본 논문에서 제안하는 모델의 Precision, Recall, F1-score 결과를 식 (15) ~ (17)과 같이 측정한다.

$$(\mathrm{Pr}ecision) = \frac{(TP)}{(TP + FP)} \tag{15}$$

$$(Recall) = \frac{(TP)}{(TP + FN)} \tag{16}$$

$$(F1 - score) = \frac{2}{(\frac{1}{\mathrm{Pr}ecision} + \frac{1}{Recall})} \tag{17}$$

4.3 실험 결과

본 논문에서 제안한 모델의 정확도는 93.5%으로 기존 제안되었던 모델들의 정확도보다 최대 30% 높은 것을 확인하였다.

Table. 5.에서 볼 수 있듯이, 본 논문에서 제안된 딥페이크 탐지 모델이 다른 모델보다 좋은 성능을 보임을 알 수 있다.

Table. 6.에서 볼 수 있듯이, XceptionNet과 LSTM을 연결한 딥페이크 탐지 모델의 AUC 값이 XceptionNet과 CLSTM을 연결한 딥페이크 탐지 모델의 AUC 값보다 0.9% 더 높게 계산되었다. AUC 값만으로 성능을 비교하였을 때, XceptionNet과 LSTM을 연결한 모델의 성능이 높게 측정된 것

Table 5. The comparison of different methods(AUC)

Model	Celeb-DF(%)
Xception[3]	67.6
Xception-raw[20]	48.2
Xception-c23[20]	65.3
Xception-c40[20]	65.5
Xception+Reg.[3]	71.2
Xception+LSTM	89.7
Xception+CLSTM	88.6
our method	**98.9**

Table 6. The comparison of different methods(Accuracy)

Model	Accuracy(%)
Xception+LSTM	69.5
Xception+CLSTM	85.0
our method	**93.5**

으로 보이지만 정확도(Accuracy)를 비교하였을 때, XceptionNet과 LSTM을 연결한 탐지 모델의 정확도는 69.5%, XceptionNet과 CLSTM을 연결한 딥페이크 탐지 모델의 정확도는 85.0%으로 XceptionNet과 LSTM을 연결한 딥페이크 탐지 모델보다 XceptionNet가 CLSTM을 탐지한 딥페이크 탐지 모델의 정확도가 15.5% 높았다. 정확도를 비교하였을 때 CLSTM을 딥페이크 탐지에 사용하는 것이 더 좋은 성능을 나타냄을 알 수 있다.

Fig. 8.은 Confusion matrix으로 본 논문에서 제안하는 모델이 데이터 셋을 판별한 결과를 표현한 그림이다. Fig. 7.의 결과를 통해 93.5%의 정확도가 계산되었다. 본 논문에서 제안한 딥페이크 탐지 모델은 동영상의 프레임 간 연결 관계를 학습하였기

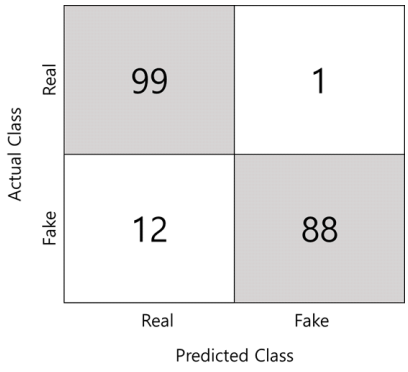


Fig. 8. Visualization of the confusion matrix

에 높은 정확도를 보였다.

Table. 7.에서 볼 수 있듯이, 본 논문에서 제안된 딥페이크 탐지 모델의 Precision은 98.9%, Recall은 88.06%, F1-score는 93.1%로 계산되었다.

Fig. 9.은 본 논문에서 제안한 모델이 각 프레임의 진위를 판별할 때 가중치를 준 부분을 Grad-CAM을 이용하여 표현한 것이다. Fig. 9.에서 가중치가 증가한 부분을 확인하면 주로 얼굴 주변 또는 입, 눈 주변에서 강조된 것을 확인할 수 있다. 딥페이크 영상 생성 시, 얼굴 주변과 입에서 불연속적인 특징이 나타나는 것을 Fig. 9.을 통해 알 수 있다. 본 논문에서 제안하는 모델을 통해 프레임간 발생하는 불연속적인 특징을 탐지할 수 있는 것을 확인하였다.

Table 7. Evaluation of our method

Method	Precision	Recall	F1-score	Acc.
our method	98.9%	88.0%	93.1%	93.5%

Fig. 9. Grad-CAM visualization results

V. 결 론

본 논문에서는 인접한 프레임의 불연속적인 특징을 시간의 순방향뿐만 아니라 역방향으로도 학습하는 딥페이크 탐지 모델을 제안하였다. 기존 합성곱 신경망과 LSTM 셀을 단순하게 연결한 모델보다 Convolutional LSTM을 이용하면 딥페이크 탐지 시에 높은 정확도를 나타내는 것을 확인하였다.

본 논문에서 제안하는 방법의 한계점은 다음과 같다. 첫째, 인접한 프레임 사이의 특징을 충분히 학습할 수 있는 프레임 개수를 알 수 없다. 둘째, 얼굴 전체를 학습하기 때문에 적절한 하이퍼파라미터(Hyperparameter)를 설정하지 않은 경우, 어텐션 모듈이 성능을 내지 못하는 한계점이 있다. 즉, 본 논문에서 제안하는 방법은 알맞은 파라미터를 설정하지

못하면 오버헤드가 크게 발생하여 기존 모델보다 성능이 떨어질 수 있다.

본 논문에서 제안하는 모델의 한계점은 프레임의 개수를 증가, 감소시키며 실험을 진행하여 학습에 적절한 프레임 개수를 탐색함으로써 개선 가능할 것으로 예상된다. 또한 본 논문에서 제안하는 방법을 통해 검출되는 불균형이 발생하는 위치를 특정하여 학습에 이용하면 하이퍼파라미터로 인한 오버헤드를 줄여 성능을 개선할 수 있을 것으로 보인다.

References

[1] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," arXiv preprint arXiv:1910.08854, 2019.

[2] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," arXiv preprint arXiv:1802.10171, 2018.

[3] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[4] W.S. Hu, H.C. Li, L. Pan, W. Li, R. Tao, and Q. Du, "Feature extraction and classification based on spatial-spectral convlstm neural network for hyperspectral images," arXiv preprint arXiv:1905.03577, 2019.

[5] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv preprint arXiv:1811.00656, 2018.

[6] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," arXiv preprint arXiv:1812.08685, 2018.

[7] T.T. Nguyen et al., "Deep Learning for Deepfakes Creation and Detection," arXiv preprints, arXiv:1909.11573, 2019.

[8] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," Remote Sens, vol. 9, no. 12, p. 1330, 2017.

[9] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," arXiv:1801.07698, 2018.

[10] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[11] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, "Bidirectional convolutional LSTM for the detection of violence in videos," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11130 LNCS, pp. 280 - 295, 2019.

[12] S. Woo, J. Park, J. Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," In Proceedings of the European Conference on Computer Vision (ECCV), pages 3 - 19, 2018.

[13] Y. Li, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in IEEE Conf. on Computer Vision and Patten Recognition (CVPR), Seattle, WA, United States, 2020.

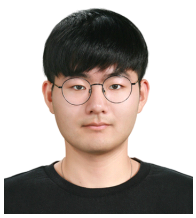
[14] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," In Neural Information Processing Systems, 2015.

[15] O.M. Parkhi, A. Vedaldi, A. Zisserma

n, “Deep face recognition.” In Proceedings of the British Machine Vision, vol. 1, no. 3, p. 6, 2015.
[16] I. Amerini, L. Galteri, R. Caldelli, and A. Bimbo, “Deepfake Video Detection through Optical Flow based CNN,” in Proc. IEEE/CVF International Conference on Computer Vision, 2019.
[17] D. Guera and EJ. Delp, “Deepfake video detection using recurrent neural networks,” In AVSS, 2018.
[18] M. Tan and Q.V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” In International Conference on Machine Learning, 2019.
[19] D.A. Pitaloka, A. Wulandari, T. Basaruddin, and D.Y. Liliana, “Enhancing cnn with preprocessing stage in automatic emotion recognition,” Procedia Computer Science, vol. 116, pp. 523-529, 2017.
[20] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. “Faceforensics++: Learning to detect manipulated facial images.” arXiv preprint arXiv:1901.08971, 2019.
[21] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, “On the Detection of Digital Face Manipulation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[22] C. Ding and D. Tao, “Robust face recognition via multimodal deep face representation,” IEEE TMM, 17(11):2049-2058, 2015.
[23] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” In CVPR, 2017.
[24] S. Suwajanakorn, S.M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing Obama: learning lip sync from audio,” ACM Transactions on Graphics (TOG), 36(4), 2017.
[25] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” In CVPR, 2014.
[26] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. OrtegaGarcia, “DeepFakes and beyond: A survey of face manipulation and fake detection,” Inf. Fusion, 2020.
[27] A. Singh, A. S. Saimbhi, N. Singh and M. Mittal, ″DeepFake Video Detection: A Time-Distributed Approach,″ SN Computer Science, 2020.
[28] K. Dale, K. Sunkavalli, MK. Johnson, D. Vlasic, W. Matusik, H. Pfster, “Video face replacement,” In Proceedings of the 2011 SIGGRAPH Asia conference. 2011. p. 1-10
[29] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and Matthias Nießner, “Face2Face: Real-time face capture and reenactment of RGB videos,” In CVPR, 2016.
[30] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 1251-8.
[31] C. Szegedy, V. Vanhoucke, S. Iofe, J. Shlens, Z. Wojna, “Rethinking the inception architecture for computer vision,” In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 2818-26.
[32] J. Hu, L. Shen, G. Sun, “Squeeze-and-excitation networks,” In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7132-41.

〔33〕 K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770 - 8.
〔34〕 DP. Kingma and M. Welling, "Auto-encoding variational bayes," In ICLR, 2014.
〔35〕 D.J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," In ICML, 2014.
〔36〕 I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," In NeurIPS, 2014.
〔37〕 ZAO.https://apps.apple.com/cn/app/zao/id1465199127. Accessed: 2019-09-16.
〔38〕 FakeApp.https://www.malavida.com/en/soft/fakeapp/, Acessed Nov 4, 2019.
〔39〕 Ajder, H, Patrini, G, Cavalli, F, et al. (2019) The state of DeepFakes: landscape, threats, and impact. Deeptrace Labs, September. Available at: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf
〔40〕 Min-seo Kim and Jong-sub Moon, "Speaker Verification Model Using Short-Time Fourier Transform and Recurrent Neural Network," Jonornal of The Korea Institute of information Security & Cryptology, 29(6), pp. 1393-1401, Feb. 2019.
〔41〕 S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," IEEE Access, vol. 6, 2018.
〔42〕 V.J. Reddi et al., "Mlperf inference benchmark," in arXiv preprint arXiv:1911.02549, 2019.

〈저자소개〉



이 대 현 (Daehyeon Lee) 학생회원
2019년 2월: 고려대학교 전자 및 정보공학과 학사
2019년 3월: 고려대학교 정보보호학과 석사 과정
〈관심분야〉 정보보호, 전자공학, 인공지능, 빅데이터, 사물인터넷

문 종 섭 (Jongsub Moon) 종신회원
1981년 2월: 서울대학교 계산통계학과 학사
1983년 2월: 서울대학교 계산통계학과 석사
1991년 2월: Illinois Institute of Technology 전산학과 박사
1993년 3월~현재: 고려대학교 전자 및 정보공학부 교수
2001년 2월~현재: 고려대학교 정보보호대학원 겸임교수
〈관심분야〉 정보보호, 운영체제, 침입탐지