딥페이크 동영상 판별 딥러닝 네트워크 모델 동작 설명 및 딥페이크 방법에 따른 시각 설명 연구

박하영 김귀식 조충상

한국전자기술연구원

hyformal@keti.re.kr, specialre@keti.re.kr, and ideafisher@keti.re.kr

A study to explain the behavior of deepfake detection and analyze network behavior according to deepfake methods using an explainable AI scheme

Hayoung Park, Guisik Kim, and Choongsang Cho

Korea Electronics Technology Institute

요 약

딥러닝 기반의 시각지능 기술은 여러 분야에서 활용되고 있으며, 합성 동영상을 만드는 주요 기술로 폭넓게 활용되고 있다. 동영상 합성 기술은 컨텐츠 제작에서 긍정적인 목적으로 활용되지만 일부 부정적인 목적으로 제작된 합성 동영상이 유통되면서 인공지능 기반의 합성 동영상 판별의 필요성이 대두되고 있다. 하지만, 딥러닝 기반 네트워크 모델은 높은 성능을 보이지만 동작을 이해하기 어려운 문제점 있기 때문에 모델들의 신뢰성을 확보하기 위한 딥러닝 모델 동작을 설명하는 연구들이 활발히 진행되고 있다. 논문에서는 동영상에서 동일 인물의 다중 얼굴 정보기반의 합성 동영상 판별 엔진의 동작을 분석하는 연구를 수행한다. 이를 위해서 기존에 연구된 시각 지능 설명 방법을 기반으로 합성 동영상 판별엔진의 결과에 대한 입력 기여도(attribution)를 분석하는 연구를 수행하며, 또한, 합성 동영상 생성 방법에 따른 입력 기여도를 분석하는 연구를 수행한다. 이를 통해 합성 동영상의 개별 입력에 대한 기여도 맵과 합성 방법에 따른 통계적인 기여도 맵을 기반으로 판별 엔진을 동작을 분석하고 딥러닝 엔진의 동작을 설명하게 된다.

1. 서 론

딥러닝 기반의 시각지능 기술은 최근 높은 성능을 보이며 자동차, 로봇, 의료, 보안, 생산 등에 폭넓은 활용되고 있다. 하지만, 얼굴 부분에 집중된 합성 동영상 생성 기술은 컨텐츠 분야에서 유용하게 활용되고 있지만, 일부 부정적인 목적으로 합성되어 유통되는 동영상은 사람이 판별하기 어려운 단계에 접근하면서 사회적 문제가 되고 있다. 이렇게 얼굴 부분의 합성에 집중된 생성 기술은 딥페이크 기술로 언급되고 있으며, 유통 및 공유되고 있는 동영상의 딥페이크를 통해서 합성된 여부를 판별하는 것이 기술의 부정적인 부분을 제거하기 위해 중요도가 높아지고 있다. 또한, 딥러닝 기반 네트워크 모델은 많이 활용되지만 동작을 설명하기 어려운 문제점이 있기 때문에 딥러닝 결과 및 판단 정보의 신뢰성을 높이기 위한 시각 Explainable AI (XAI)연구들이 활발히 진행되고 있다.

본 연구에서는 시각지능에 대한 XAI에서 활발히 연구된 방법을 활용하여 동영상에 대한 합성 여부를 판별하는 딥러닝 네트워크 모델의 동작에 대한 설명 및 분석 수행한다. 이를 위해서 최신의 판별 딥러닝 모델을 Backbone으로 적용한 다중 얼굴 이미지 기반의 합성동영상 판별 엔진을 학습하고, 학습된 딥러닝 네트워크에 XAI 방법을 적용하여 합성 여부 판별에 대한 입력 attribution 맵 추정하는 분석을 수행한다. 또한 딥페이크 동영상 생성 기법에 따른 특성을 분석하기 위해서 딥페이크 생성 방법별로 획득된 입력 attribution(기여도)을 중첩하여 딥페이크 기법 특성이 반영된 분석을 함께 수행한다.

본 연구의 내용을 통해서 시각 지능을 활용하는 응용 기술의 한 부분에서 시각 XAI 기술을 활용하여 특성을 분석하는 과정 및 내용에 대해서 좀 더 자세히 설명한다.

2. 딥페이크 기술 및 판별

딥러닝 기반의 이미지 생성 및 변환 기술을 기반으로 얼굴에 집중된 합성 영상 생성 방법은 딥페이크 영상-동영상 생성 기술이라고 언급된다. 최근 시각 데이터 관련 auto-encoder [1], Generative adversarial network(GAN) [2], diffusion model [3] 등의 생성을 위한 딥러닝 모델이 연구되면서 딥페이크 동영상의 품질도 향상되고 있다.

본 연구에서는 다양한 종류의 딥페이크 방법으로 생성된 동영상을 판별하기 위해서 그림 1과 같이 동일

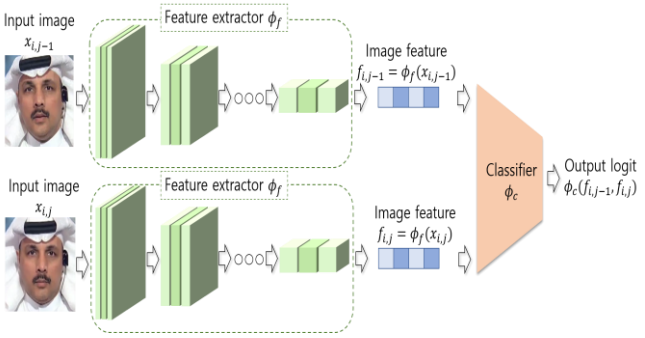

그림 1. 동일 인물 얼굴 이미지셋 기반 딥페이크 탐지/판별 모델

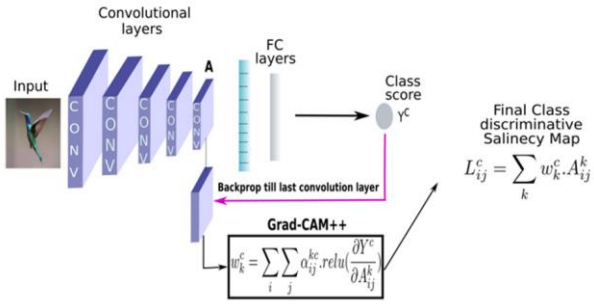

그림 2. Grad-CAM++의 동작 [8]

인물의 연속적인 여러 얼굴 이미지로부터 해당 얼굴 이미지에 대한 특징 벡터를 추출하고, 여러 이미지의 특징 벡터를 하나의 벡터로 통합하여 딥페이크 이미지 여부를 판단하는 과정을 수행한다. 동영상에서 동일 인물을 추적하며 얼굴 영역을 이미지를 획득하기 위해서 본 연구에서는 입력된 동영상을 사전에 학습된 딥러닝 기반의 사람 추적과 얼굴 검출 엔진을 활용하여 동일 인물의 위치를 추적하고 해당 인물의 얼굴 영역을 검출하여 동일 인물의 얼굴 이미지셋 기반으로 입력된 데이터가 딥페이크 기술이 적용된 이미지인지 여부를 판별하는 과정을 수행한다.

3. 모델 출력 분석 방법

본 연구에서 XAI 기반의 시각 설명 기술을 딥페이크 동영상 판별 기술의 동작 분석 및 딥페이크 동영상 생성 방법의 특성을 분석하는 연구를 수행하기 위해서 시각 지능의 판별 기술을 위해 연구된 XAI을 응용한다. 기존 시각 지능의 분류 네트워크의 동작을 설명하기 위한 연구로 Class Activation Mapping(CAM) 기반의

연구[6, 7, 8] 와 네트워크의 내부 도출 정보 없이 동작하는 RISE [9]등이 많이 사용되고 있다. 본 연구에서는 딥페이크 판단 엔진이 동영상 데이터를 입력으로 활용하기 때문에 네트워크 동작 설명 분석에 활용하는 방법의 복잡도가 낮으면서 안정적인 성능을 갖는 그림 2와 같은 Grad-CAM++[8]을 적용한다. 이를 위해서 본 연구에서 딥페이크 판별 Feature 추출 네트워크의 마지막 레이어에서 획득 가능한 activation 맵과 손실 연산 후 수행되는 역전파를 통해 획득되는 Gradient 정보를 함께 이용하여 기존에 정립된 Grad-CAM++의 수식을 통해서 해당 결과 판별에 대한 판별 네트워크의 입력 attribution 맵을 획득하게 된다.

4. 실험 및 분석

4.1 딥페이크 판별 엔진 학습

본 연구에서 딥페이크 판별 모델을 위해서 특징 추출 네트워크를 위해서 EfficientNet [5]을 적용했으며, 딥페이크 판별 네트워크는 H.264 비디오 코덱의 압축 팩터 23% 조건으로 저장된 faceforensics++[4] 데이터셋으로 학습이 진행되었다. 학습에 사용된 faceforensics++ 데이터 셋은 1000개의 원본 동영상과 5가지 방법으로 (Deepfakes, Face2Face, FaceSwap, FaceShifter, and NeuralTextures) 생성된 동영상으로 구성된다. 딥페이크 생성 방법 중 두 이미지에 포함된 얼굴을 교환하는 방법(Deepfakes, FaceSwap, FaceShifter)이 활용되었고, 소스 이미지에 포함된 얼굴의 표정과 동작을 목표 얼굴 이미지에 반영(face reenactment) 방법으로 Face2Face과 NeuralTextures 방법이 활용되었다.

Deepfakes 방법은 두 개의 Auto-encoder를 사용하여 소스 이미지에 포함된 얼굴을 학습한 모델을 목표 이미지에 포함된 얼굴 이미지에 적용함으로써 얼굴 교환을 수행된다. FaceSwap 방법은 고전 그래픽스 기술을 활용하여 얼굴 특징점을 추출하고 얼굴 영역을 3차원 템플릿 모델 맞추는 방식 수행하게 되며, FaceShifter는 GAN 기반 네트워크로 소스와 목표 얼굴 특징을 결합하여 소스 얼굴 형태에 가까운 합성된 얼굴을 자연스럽게 생성한 후 딥러닝 기반의 facial occlusion 처리하는 방식으로 얼굴 교환을 수행한다. Face reenactment 계열의 방법인 Face2Face는 딥러닝 기술 없이 얼굴 특징 정렬과 통계적 regularization을 포함 최적화 방법으로 딥페이크 이미지 생성이 이루어진다. 그리고, NeuralTextures 방법의 경우는 딥러닝 기반 neural texture 정보를 추출하고 뉴런 렌더링을 통해서 모사된 얼굴 이미지가 획득된다.

4.2 딥페이크 판별 네트워크 동작 시각 설명 및 분석

딥페이크 판단 네트워크의 결과를 기반으로 동작을 분석하기 위해서 Grad-CAM++를 기반으로 입력의 영향도(attribution 맵)를 분석하게 된다. 얼굴 이미지에서 해당 위치의 attribution 맵 값이 큰 경우 판별에 높은 영향을 갖는 것을 의미하며, 작은 경우는 낮은 영향도를 갖는 것을 나타낸다.

그림 3은 faceforensics++ 데이터셋에 포함된 비디오 데이터에서 추출된 얼굴 영역 이미지를 기반으로 딥페이크 판별 네트워크에 적용하고 얻은

그림 3. Grad-CAM++을 통해 얻은 입력 attribution 맵과 입력 이미지 중첩한 결과

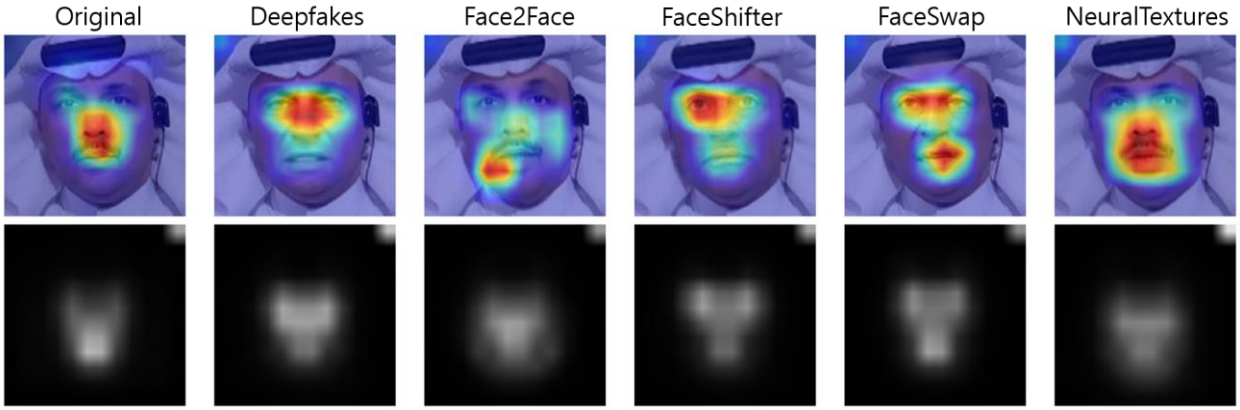

그림 4. 딥페이크 방식에 따른 입력 attribution 맵 및 attribution 누적한 결과

결과에 대한 정보를 Grad-CAM++ 활용하여 네트워크 결과에 대한 입력 attribution에 대한 맵 정보를 입력 얼굴 이미지에 중첩하여 보여준 결과이다. 앞에서 설명했듯이 딥페이크 시각데이터 생성 기술은 사람의 얼굴 영역에 중점을 두고 있으며, 특히 사람 얼굴에 대한 특징 결정에 높은 영향도를 갖는 눈/코/입 영역의 입력 attribution 맵 이 큰 값을 갖는 것을 확인할 수 있다. 이러한 결과는 딥페이크 판별 네트워크가 목적에 맞는 동작을 하고 있는 것으로 평가할 수 있다.

본 연구에서 그림 4와 같이 딥페이크 생성 방법에 따른 특성을 분석하기 위해서 동일한 소스 이미지 기반 5가지 딥페이크 방법으로 합성된 이미지에 대한 판별과 5가지 딥페이크 방법의 합성된 다양한 이미지에 대한 판별 결과를 통계적으로 분석하는 과정을 수행한다. 소스와 목표 이미지 포함된 얼굴 부분을 교환하는 방식의 합성 방식은 사람의 얼굴 특징에 중요도가 높은 눈, 코, 입 주변에 대한 판별에 대한 공헌도가 높은 것을 확인 할 수 있다. 또한 얼굴의 표정과 동작을 모사시키는 Face reenactment 계열의 방법은 얼굴의 넓은 영역에 대한 입력의 공헌도가 분포하는 것을 확인할 수 있다.

통계적인 딥페이크 방법별 특성을 분석하기 위해서, 테스트 비디오 셋을 입력으로 딥페이크 방법별 획득된 입력 attribution 맵을 누적하고, 평균을 구한 후 정규화한 값을 이미지 형태로 그림 4의 하단과 같이 출력하였다. 딥페이크 방법에 따른 통계적 판단 네트워크 동작 설명 정보를 보면, GAN을 활용한 FaceSwap의 결과와 FaceSifter, Deepfakes의 결과가 눈과 입 주변으로 높은 기여도를 갖는 것을 알 수 있다. 또한 Auto-Encoder가 사용된 Deepfakes 방법의 결과를 보면 다른 방법 대비 양쪽 눈사이의 정보가 공헌도가 높은 것을 알 수 있다.

Face reenactment 계열의 딥페이크 방법의 통계적인 특성 분석 결과를 보면 코와 입을 중심으로 주변부에서 기여도를 갖는 것을 확인할 수 있다. 이러한 것은 딥페이크 방법으로 생성된 이미지를 보면 소스 이미지의 눈과 딥페이크 방법으로 생성된 눈이 유사성을 갖고 있으며 입 주변의 움직임이 모사 되면서 코와 입주변의 입력값이 판별에 높은 공헌도를 갖는 것을 추정할 수 있다.

이러한 특성 분석은 딥페이크 동영상을 생성하는 방법이 어느 부분을 집중해서 합성된 얼굴 이미지를 만들었는지 확인할 수 있다. 또한, 딥페이크 생성 방법이 의도한 특성에 맞게 딥페이크 동영상을 만들었지 혹은 새로운 딥페이크 생성 방법인지 검토하는 유용한 방법이 될 것으로 보인다.

5. 결론

본 논문에서는 얼굴 영역에 집중된 합성 방법인 딥페이크 기술이 적용되어 변형된 동영상을 찾아내기 위한 동영상 판별 기술에 대해서 언급하고, 시각 지능 설명 기술로 많이 활용되고 있는 설명가능 인공지능

기술을 접목하여 딥페이크 동영상 판별 네트워크의 동작을 설명하는 과정을 수행했다. 딥페이크를 판별하는 시각 지능을 분석하기 위해서 판별에 얼굴 영역 부분이 높은 영향도를 갖는지 확인하고, 딥페이크 생성 방법 특성에 따른 분석을 진행했다. 동일한 원본 동영상 기반 5가지 딥페이크 방법으로 생성된 이미지의 얼굴로 판별된 결과 딥페이크 방법별 특성이 반영된 입력 공헌도를 갖는 것을 확인했다. 또한, 판별 결과에 대한 설명 정보를 통계적으로 분석한 결과 딥페이크 생성 방법의 특성이 잘 나타내는 입력 attribution 맵을 보이는 것을 확인했다.

본 연구에서 시각 설명가능 기술로 시각 지능 기술에 적용하여 네트워크의 동작 및 분석을 위한 정보를 제공하는 과정을 보였다. 이러한 연구 과정은 시각 설명가능 기술을 시각 응용 기술에 활용하는 방법 및 과정을 보여준 것으로 다양한 시각 응용 기술에 설명가능 기술을 활용하는 안내 연구로 의미를 갖는다.

감사의글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00984)

참고문헌

[1] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” Int’l Conf. on Learning Representations (ICLR), 2014.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” Communications of the ACM, vol. 63, no. 11, pp. 139-144, 2020.

[3] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” Advanced in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840-6851, 2020.

[4] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: learning to detect manipulated facial images,” Int’l Conf. on Computer Vision (ICCV), 2019.

[5] M. Tan and Q. Le, “Efficientnet: rethinking model scaling for convolutional neural networks,” Proc. of the 36th Int’l Conf. on machine learning (PMLR), vol. 97, pp. 6105-6114, 2019.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2921-2929, 2016.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: visual explanations from deep networks via gradient-based localization,” Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 618-626, 2017.

[8] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” IEEE Winter Conf. on Applications of Computer Vision (WACV), pp. 839-847, 2018.

[9] V. Petsiuk, A. Das, and K. Saenko, “RISE: randomized input sampling for explanation of black-box models,” in Proc. of the British Machine Vision Conference (BMVC), 2018.