동영상에서 음성과 이미지 데이터를 이용한 진술의 거짓말 탐지

양지석*, 진예섬*, 이승우**, 원일용*
*호서전문학교 사이버해킹보안과
**호서전문학교 컴퓨터공학과
jiseok4404@naver.com, jejjnd@naver.com, as_thtls@naver.com, clccclcc@shoseo.ac.kr

Lie Detection of statements using voice and image data in the video

Ji-Seok Yang*, Ye-Seom Jin*, Seoung-Woo Lee**,Ill-young Weon*
*Dept. of Cyber Security, Hoseo technical College
**Dept. of Computer Engineering, Hoseo technical College

요 약

경찰 수사에서 진술의 진실 여부를 인공지능 기법을 이용하여 판단하는 연구는 인적, 물적 자원의 낭비를 줄일 수 있다. 우리는 진술 동영상에서 이미지, 음성 데이터를 각각 추출하여 동시에 고려해 진술의 진실 여부를 자동으로 판단하는 시스템을 제안하였다. 실험을 통해 제안된 시스템이 유의미함을 알 수 있었다.

1. 서론

경찰청이 발표하는 통계에 의하면 거짓말 탐지 검사 의뢰 건수가 2014 년에는 9,845 건, 2018 년에는 11,256 건으로 점차 증가세를 보이고 있다. 오늘날 과학기술의 발달로 지능화된 현대 범죄에 대해 더욱 효율적이고 과학적인 수사기법의 필요성이 커지고 있다. 이러한 과학 수사 기법의 하나가 진술에 대한 진실 여부를 판단하는 거짓말탐지기를 이용하는 것이다[1].

거짓말탐지기를 이용하는 방식은 거짓말 탐지에 드는 인적, 물적 낭비가 크다. 따라서 인공지능 기법을 이용한 자동 거짓말 탐지에 관한 연구가 진행되고 있다[1]. 최근 진술하는 음성 또는 이미지를 기반으로 거짓말을 탐지하는 연구가 진행되고 있지만, 좋은 성과를 얻은 연구는 알려져 있지 않다[2, 3].

본 논문은 진술인의 동영상을 인공지능 기법을 사용하여 학습하고, 이것을 기반으로 진술의 진실 여부를 자동으로 판단하는 시스템에 관한 연구이다. 우리가 제안하는 시스템은 진술 동영상에서 이미지 부분과 음성 부분의 데이터를 추출하고, 이 두 데이터를 동시에 고려한 학습을 통해 진술의 진실 여부를 판단하는 시스템이다. 제안된 시스템의 유용성은 실험으로 검증하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구를 언급하였다. 3 장에서는 이미지와 음성 거짓말 탐지 모델을 제안하였다. 4 장에서는 실험 및 결과를 제시하고 5 장에서는 결론 및 앞으로 연구 방향을 언급하였다.

2. 관련 연구

2.1 영상 학습

본 논문의 동영상 학습을 위해서 얼굴 부분을 추출하는데 이때 OpenCV 와 Dlib 를 사용하였다. OpenCV 는 오픈소스 컴퓨터 비전 C 라이브러리이다. OpenCV 는 실시간으로 영상을 처리하는 요구조건을 만족시키며 상용으로 사용할 수 있는 BSD 라이선스를 가지고 있다[4].

Dlib 는 딥러닝을 포함한 다양한 머신 러닝 알고리즘 기능을 포함하고 있으며, 얼굴 인식에 쓰이는 대표적인 라이브러리 중 하나이다[5].

또한 일반적으로 모델의 크기를 증가시키면 정확도와 연산량이 증가한다. 이러한 연산 량이 증가하는 단점을 보완하기 위하여 적은 파라미터를 가진 42-layer 의 깊은 신경망이지만 VGGNet 과 비슷한 연산량을 가진 Inception-V3 를 사용하기도 한다[6].

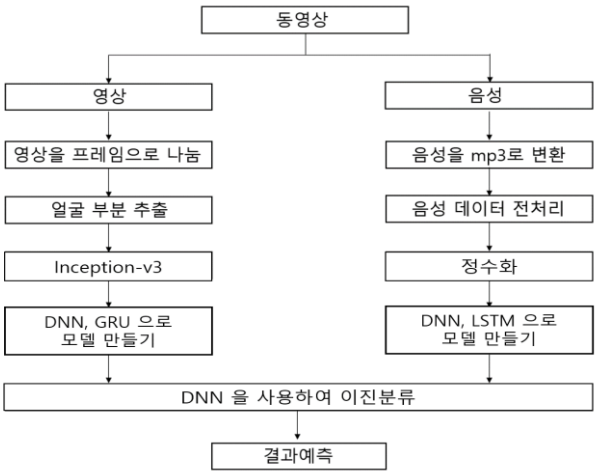
2.2 음성 학습

STFT 란 시간이 지남에 따라 변화하는 신호의 사인파 주파수와 위상 성분을 결정하는 데 사용되는 푸리에 관련 변환이다[7]. STFT 는 시간에 따라 변화하는

긴 신호를 짧은 시간 단위로 분할한 다음에 푸리에 변환을 적용하기에 결과적으로 각 시간 구간마다 어떤 주파수들이 존재하는지 알 수 있다. Librosa 라이브러리에는 STFT 함수를 제공한다.

3. 거짓말 탐지 시스템

우리가 제안하는 시스템의 구성은 아래와 같다. 학습 대상의 동영상에서 영상과 음성을 추출하여 학습하고 최종적으로 이 두 가지의 결과를 이용하여 진술의 참여부를 판단하는 구조이다.

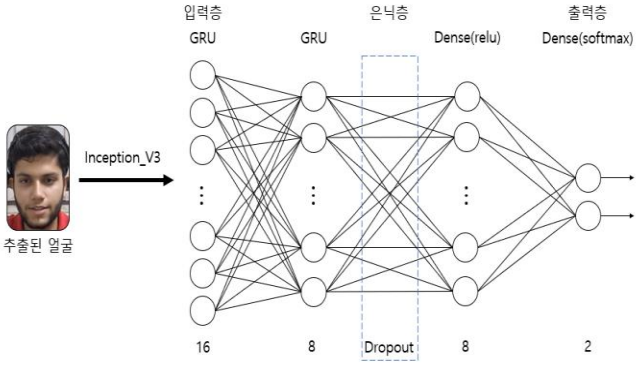


(그림 1)전체 학습 흐름도

3.1 영상 데이터 추출 및 학습

학습할 동영상에서 먼저 일정 시간 간격으로 이미지 프레임을 추출하고, 추출된 이미지에서 얼굴 영역을 2 차로 추출한다. 이렇게 추출된 얼굴 이미지는 학습을 위해 일정 크기의 동일 크기 영상으로 다시 조정된다. 그리고 조정된 이미지는 가장 먼저 Inception-V3 모델에서 한 번 학습된 후 GRU 신경망에 들어가게 된다.

추출된 이미지에서 특정 이미지 1 개에서 거짓의 징후를 판단하는 것은 어렵기 때문에 관련된 여러 개의 이미지를 동시에 고려해서 학습해야 한다. 학습기는 아래와 같이 구성하였다.

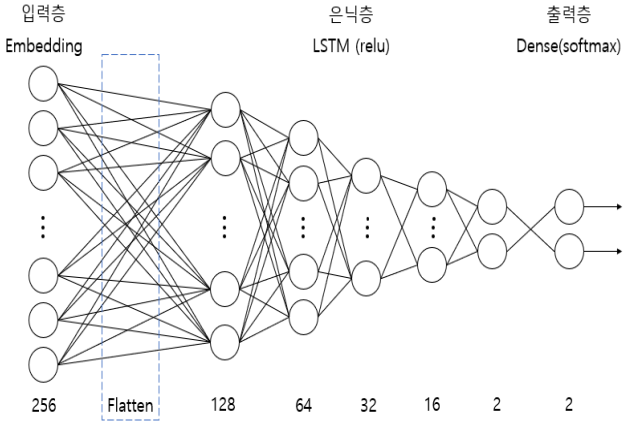


(그림 2)GRU 를 이용한 동영상 학습 구조

3.2 음성 데이터 학습

동영상에서 음성 데이터의 추출은 다음과 같다. 먼저 음성데이터를 단일 채널로 통합한 후 수치 변형 알고리즘을 이용하여 수치 데이터로 추출하는 과정이 진행된다. 이렇게 추출된 데이터는 영상처럼 연속적으로 고려해야 하는 데이터이기에 시계열을 고려한 학습 알고리즘이 필요하다.

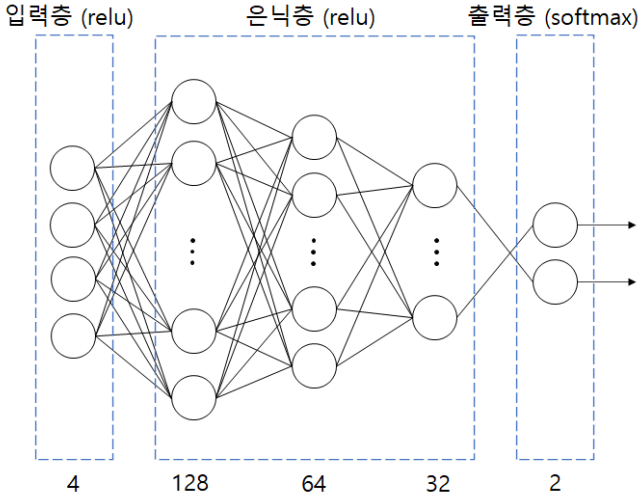
본 논문에서는 Embedding 층에서 수치 데이터를 벡터화 시킨 후 LSTM 층에서 시계열 학습을 시킨다.



(그림 3) 음성 데이터 학습을 위한 신경망 구조

3.3 영상과 음성 데이터의 앙상블(Ensemble)

영상 데이터 또는 음성 데이터 단독의 학습 모델보다는 두 개의 데이터를 동시에 고려하는 것이 더욱 효과가 높다고 가정한다. 따라서, 우리가 제안한 시스템에서는 두 가지 데이터를 동시에 고려하여 최종적으로 참, 거짓을 판단하는 구조이다. 해당 신경망에서는 간단한 Dense 층만을 사용하여 구성한다. 이 때 입력 데이터는 영상 학습의 출력 unit 과 음성 학습의 출력 unit 을 결합한 4 개의 unit 이 된다.



(그림 4) DNN 을 이용한 최종 판단 신경망 구조

4. 실험 및 결과

4.1 실험환경 및 데이터

제안 시스템의 유용성을 검증하기 위한 시스템을 구현하고 딥러닝을 통해 결과를 분석하였다. 시스템 구현은 Python, Tensorflow, Opencv 등을 사용하였다.

실험을 위한 데이터 셋은 BagofLies[8]을 사용하였는데, 이 데이터 중 162 개의 거짓말 영상과, 163 개의 진실 영상을 사용하였으며 학습과 테스트를 위하여 아래 표처럼 구성하였다.

	Train	Test
진실	114	49
거짓	113	49

이미지 및 음성 학습기의 구조는 앞장에서 설명한 구조를 사용했으며 각 구조의 하이퍼 파라미터는 반복 실험을 통해 적절하게 설정하여 사용하였다.

4.2 실험 결과

학습기의 성능 측정을 위해 순수 DNN 과, LSTM, GRU 알고리즘을 적용하여 실험하였다. 또한, 음성만으로 판단한 경우와 영상만으로 판단한 경우, 그리고 두가지를 동시에 고려해 판단한 경우로 구분하여 실험하였다. 실험 결과는 아래와 같다.

먼저 DNN 알고리즘에서 음성과 영상 모두 각각 판단했을 경우를 비교하면 유의미한 결과를 볼 수 없었지만 음성과 영상을 동시에 고려했을 때는 유의미한 결과를 볼 수 있었다.

LSTM 알고리즘에서 음성만 판단한 경우를 비교해 보면 향상된 결과를 볼 수 없었지만 영상만 판단한 경우는 정확도가 14%가 증가한 것을 확인할 수 있었다. 음성과 영상을 모두 고려한 상황은 음성 또는 영상만으로 판단한 경우보다 높은 정확도가 나왔지만 DNN 알고리즘을 사용했을 때 보다 미약하게 향상된 결과를 볼 수 있었다.

	음성	영상	음성, 영상 고려
A(DNN)	0.51	0.48	0.73
B(LSTM, GRU)	0.45	0.62	0.75

(그림 5)DNN 과 LSTM 의 실험결과

그림 6 은 음성과 영상을 동시에 고려한 경우의 실험결과이다. 모델의 안정성 검증을 위해 총 10 번 교차검증을 진행하였다.

	1	2	3	4	5	6	7	8	9	10
A	75	66	66	72	81	81	75	68	62	75
B	66	78	81	87	78	56	65	65	78	75

(그림 6)A(DNN), B(LSTM)의 교차검증 결과

총 10 번의 실험을 거친 결과, 대부분의 실험에서 60%에서 80%로 일관된 정확도를 보여 해당 실험이 유의미함을 증명할 수 있었다.

5. 결론 및 향후 과제

범죄 수사에서 진술의 진정성 여부를 탐지하는 기존 방법은 많은 시간적, 물적 자원이 필요하다. 이에 인공지능 기법을 이용하여 자동으로 진정성 여부를 판단하는 연구가 필요하다.

우리는 기존의 연구와는 다르게 진술 영상에서 영상 부분과 음성부분을 동시에 고려해서 학습하는 시스템을 제안하였다. 제안된 시스템의 핵심은 영상 데이터와 음성 데이터를 시계열 데이터로 처리한다는 것과 영상과 음성을 동시에 고려해 학습한다는 점이다.

제안된 시스템의 유용성을 위해 실험을 실시하였고, 어느 정도 유의미한 결과를 얻을 수 있었다.

향후 좀 더 많은 학습 데이터를 확보하여 알고리즘의 유용성을 검증하는 일이 필요하며, 다양한 신경망 모델을 결합한 알고리즘의 확장이 필요하다.

참고문헌

[1] 송승은, “거짓말탐지기 검사와 그 결과의 증거능력에 관한 고찰”, 성균 관대 법학연구원, 제 18 권 제 3 호, pp.535-562, 2006

[2] 강민수, 홍훈기, 구재훈. (2020). 얼굴분석기반 거짓말탐지기법 연구. 한국통신학회 학술대회논문집, 248-249.

[3] 김영명. "양방향성 장단기 기억 기술 순환신경망 모델 기반 음성 거짓말 탐지 알고리즘 개발." 국내석사학위논문 한양대학교 의생명공학전문대학원, 2020. 서울

[4] 강석원, 이순이, 박지웅."OpenCV 를 사용한 화재영상 처리."한국콘텐츠학회 종합학술대회 논문집 7.1(2009):79-82

[5] 허석렬, 김강민 and 이완직. (2021). 딥러닝 얼굴인식 기술을 활용한 방문자 출입관리 시스템 설계와 구현. 디지털융복합연구, 19(2), 245-251.

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826

[7] T. Baba, “Time-Frequency Analysis Using Short Time Fourier Transform,” Open Acoustics Journal, 2012
[8] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh and M. Vatsa, "Bag-of-Lies: A Multimodal Dataset for Deception Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 83-90, doi: 10.1109/CVPRW.2019.00016.