CNN-LSTM 모델 기반 딥 보이스 탐지

강구현○ 권민혜

숭실대학교

rkdrngus22@soongsil.ac.kr, minhae@ssu.ac.kr

CNN-LSTM model Based Deep Voice Detection

Guhyeon Kang○ Minhae Kwon

Soongsil Universitiy

요 약

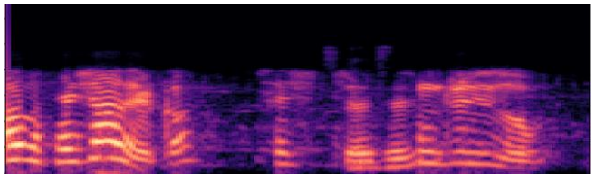
본 논문에서는 CNN, LSTM 을 함께 사용한 하이브리드 신경망(Hybrid Neural Network)을 통해 딥 보이스를 탐지하는 모델을 제안한다. 제안한 모델은 멜 스펙토그램(Mel spectrogram)으로 전처리한 음성 데이터의 공간적 특징과 시간적 패턴을 학습함으로써 음성 데이터의 복잡한 구조를 더 잘 이해하고 해석할 수 있다. 본 논문에서 제안한 모델은 오픈 소스 딥 보이스 모델인 RVC(Retreival-Based Voice Conversion)로 생성한 음성에 대한 딥 보이스 탐지를 진행한다.

1. 서 론

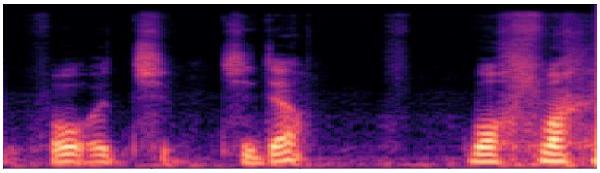
AI 기술의 발전은 일상에 많은 변화를 가져왔지만, 동시에 그 기술을 악용하는 범죄의 형태도 등장하고 있다. 특히 딥 보이스 기술을 활용한 가짜뉴스, 보이스 피싱은 심각한 사회적 문제로 받아들여지고 있다. 이에 대항하기 위하여 AI 를 이용한 예방 및 탐지 대책이 필요하다. 본 논문에서 제안하는 CNN - LSTM 하이브리드 신경망 모델은 CNN 을 통해 데이터의 공간적 특징을 학습하고 LSTM 을 통해 시간적 패턴을 학습한다. 이는 음성 데이터의 감정을 인식하는데 높은 성능을 보여준다.[1] 이를 활용하여 실제 음성과 딥 보이스 음성의 미묘한 감정 차이를 학습할 수 있다.

2.1 데이터셋 구성 및 전처리

본 논문에서 제안하는 모델은 RVC 를 사용하여 생성된 딥 보이스 음성과 진짜 음성을 이진 분류하는 모델이다. 이를 위해 정상적인 음성 파일과 RVC 를 통해 변조된 데이터가 필요하다. 학습 데이터 셋으로 AIhub 의 감정이 태깅된 자유대화(성인)[2]를 사용하였으며, 이는 전화 통화를 통해 두 사람이 대화를 나누는 형식으로 구성되어 있다. 이 중에 약 183 시간 분량의 데이터를 사용하였고 그 중 절반을 RVC 딥 보이스 음성으로 변조시켰다. 데이터 셋의 다양성 확보를 위해 10 개의 RVC 보이스 모델을 활용하여 딥 보이스 음성을



실제 음성의 멜 스펙토그램



RVC 변조 음성의 멜 스펙토그램

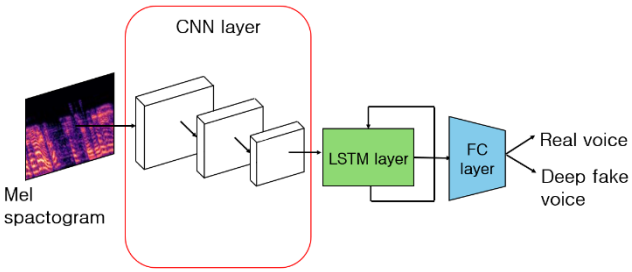


그림 1. CNN – LSTM 모델의 구조도

생성하였고, 이들 모델은 다른 데이터셋[3]과 유명인의 음성 샘플을 사용하여 학습시켰다. 최종적으로 모든 음성 파일을 2 초 간격으로 나누어 멜 스펙토그램을 통해 128 개의 특성을 추출하였다.

모델을 평가할 테스트 데이터셋은 세 개를 준비하였다. 첫 번째 테스트셋은 모델이 학습 과정에서 사용한 데이터와 동일한 분포를 가진 데이터로 구성되어 있다. 이를 통해 모델이 학습 데이터를 잘 이해하고 있는지를 평가한다. 두 번째 데이터셋은 학습 데이터의 특성을 모방하여 제작한 데이터이다. 역시 전화 통화 데이터이며 모델이 학습 데이터의 일반적인 패턴을 잘 학습했는지를 평가한다. 세 번째 데이터셋은 실제 환경에서 얻은 다양한 음성 데이터를 포함한다. 모델이 일상생활의 다양한 환경에 얼마나 효과적으로 작동하는지를 평가한다. 각각의 테스트 셋은 3400, 1352, 1666 개의 샘플 수를 가지고 있다.

1. 학습 데이터와 동일한 테스트 셋[2][3]
2. 학습 데이터를 모방하여 제작한 테스트 셋
3. 유튜브 등에서 녹음한 테스트 셋[4]

2.2 모델 구성 및 하이퍼파라미터

멜 스펙토그램으로 전처리 된 데이터들은 시간에 따른 주파수 영역의 변화를 담고 있는 2 차원 데이터이다. 이는 이미지 데이터처럼 다루어질 수 있고 CNN 은 이러한 2 차원 이미지 데이터의 특징을 잘 학습하는

<표 1> 모델의 하이퍼 파라미터

HyperParameter	Value
Number of input	321,536
Input size	(128,188)
Batch size	512
Epoch	20
Learning rate	0.0001
Loss Function	CrossEntropyLoss

<표 2> 모델 구성

Layer Type	Output Size	Kernel Size
Conv2d	(32,128,188)	(3,3)
ELU	(32,128,188)	-
BatchNorm2d	(32,128,188)	-
Maxpool2d	(32,64,94)	(2,2)
Conv2d	(64,64,94)	(3,3)
ELU	(64,64,94)	-
BatchNorm2d	(64,64,94)	-
Maxpool2d	(64,32,47)	(2,2)
Conv2d	(128,32,47)	(3,3)
ELU	(128,32,47)	-
BatchNorm2d	(128,32,47)	-
Maxpool2d	(128,32,47)	(2,2)
LSTM	(512)	-
Dense	(256)	-
ELU	(256)	-
Dropout	(256)	-
Output	(2)	-

딥러닝 네트워크 중 하나이다. 한 개의 CNN 모듈은 컨볼루션 레이어, ELU 활성화 함수, Batch Normalization, Max Pooling 레이어 순으로 구성하였다. 또한 음성 데이터는 시퀀스를 가지는 시계열 데이터이다. LSTM(Long Short-Term Memory) 네트워크는 RNN 의 한 유형으로 긴 시퀀스 데이터에서 장기간에 걸친 의존성을 학습하는데 높은 성능을 보인다. 본 논문의 모델에서는 2 개의 Layer 를 가지고 정방향과 역방향의 패턴을 모두 학습하는 Bidirectional LSTM 을 사용하였다. LSTM 에서 출력된 특징 벡터의 마지막 시퀀스 출력에 대해 완전 연결 계층을 통과시켜 최종 출력을 얻는다.

2.3 모델 평가

그림 2 는 모델의 Learning Curve 그래프이다. 적은 epoch 수에도 최종 값에 수렴하였으며 과적합 없이

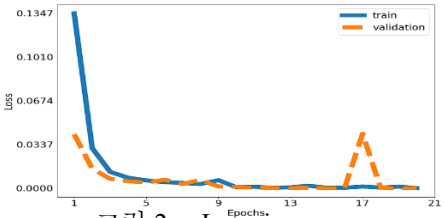


그림 2. Learning curve

훈련을 종료하였다. train 데이터셋에 대해서는 99.82%,

3.1 테스트 셋 1 3.2 테스트 셋 3.3 테스트 셋 3

테스트 셋	accuracy	Precision	Recall	F1score	Fallout
1	99.56%	99.1%	100%	99.6%	0.9%
2	91.42%	89.5%	90.7%	90.1%	8.1%
3	58.76%	28.2%	72.5%	40.6%	44.6%

<표 2> 모델 성능 평가 결과

validation 데이터 셋에 대해서는 99.63%의 정확도를 기록하였다.

그림 3 과 <표 3>은 각각의 테스트 셋에 대한 오차행렬과 모델 성능 평가 결과이다. 이를 보면 알 수 있듯이 모델은 첫 번째와 두 번째 테스트 셋에 대해서는 비교적 정확한 판단을 하고 있지만 세 번째 데이터 셋에 대해서는 정상적으로 동작하고 있지 않다. 세 번째 테스트 셋에 대한 accuracy 와 precision 이 각각 58.76%, 28.2%로 모델이 실제 음성을 RVC 변조 음성으로 판단하고 있기 때문이다.

3. 결론 및 발전방향

본 논문에서 제안한 CNN-LS TM 모델은 학습 데이터 및 그와 유사한 데이터에 대해 높은 성능을 보여주며, 이는 통화 데이터에 대한 우수한 인식률을 의미한다. 하지만 유튜브 등에서 녹음한 보다 일반적인 데이터에 대해서는 정상적으로 동작하지 않는다. 이는 모델이 음성 데이터의 보편적인 패턴을 학습하지 못하였음을 시사하며, 이에 대한 원인으로 학습 데이터 셋의 다양성 부족을 지적할 수 있다. 학습 데이터 셋은 통화 음성만으로 구성되어 있으며 모델은 이러한 데이터의 특징에 과적합 되어있다고 판단된다.

이러한 문제점을 개선하기 위하여 우선적으로 학습 데이터셋의 다양성을 확대하여야 한다. 우리가 딥 보이스를 가장 많이 접할 수 있는 환경은 유튜브를 비롯한 동영상 플랫폼이다. 따라서 해당 플랫폼에서 다양한 배경, 성별, 연령대, 방언을 포함하는 음성 데이터를 수집하고 이를 학습 데이터셋에 통합하여야 한다. 더 나아가 RVC 뿐만 아니라 일반적인 딥 보이스에 대한 탐지모델을 구현하기 위하여 다양한 알고리즘으로 생성된 딥 보이스의 특징을 연구하고 데이터를 수집하여야 한다. 또한 딥 보이스 탐지 모델의 전체적인 성능 향상을 위하여 CNN-LSTM 모델에 국한되지 않고 transformers 나 GAN 등 다양한 딥러닝 네트워크를 실험해 보아야 한다. 이를 통해 더욱 개선된 딥 보이스 탐지 모델을 구현할 예정이다.

참 고 문 헌

[1] 윤상혁, 전다윤, 박능수 CNN-LSTM 모델 기반 음성 감정인식. “한국정보처리학회 학술대회논문집 28 권 2 호 939-941,”

[2] Alhub 데이터셋 감정이 태깅 된 자유대화(성인)

[3] Alhub 데이터셋 화자 인식용 음성 데이터

[4] Alhub 데이터셋 동영상 콘텐츠 하이라이트 데이터