음성 위조 탐지를 위한 앙상블 모형 연구

강태인[1] , 곽일엽[2]

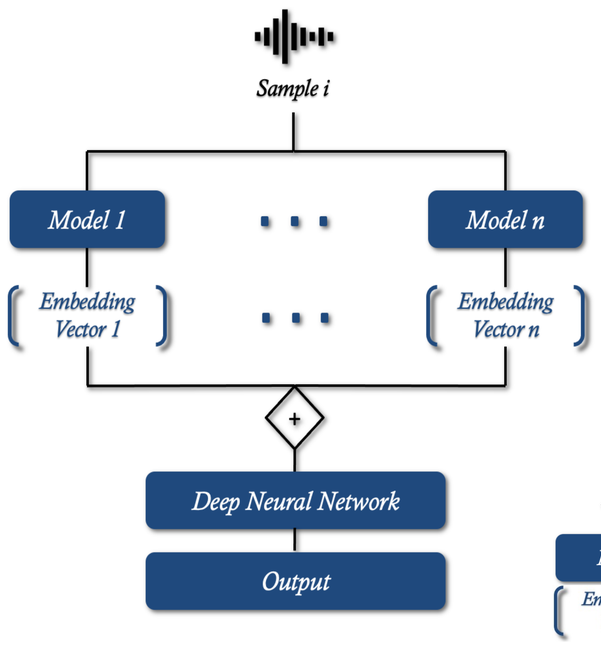
[1] (06974) 서울시 동작구 흑석로 84, 중앙대학교 일반대학원 통계학과. E-mail: xodls4179@cau.ac.kr

[2] 교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 경영경제대학 응용통계학과. E-mail: ikwak2@cau.ac.kr

Introduction

- 음성 위조 탐지는 음성을 여러가지 방법으로 재생산한 위조 음성과 원본 음성을 구별하는 문제이다.
- 음성 비서와 같이 실생활에서 사용자의 음성으로 여러가지를 수행하거나 사용자를 인증하는 기기가 증가함에 따라 실제 음성과 음성을 재생산한 위조 음성 (예를 들어 TV에서 나오는 음성 혹은 해킹을 시도하기 위해 만들어진 음성) 을 구분 할 필요성이 있다.
- 제시하는 앙상블 모형은 모형 성능 척도인 EER(Equal Error Rate)를 적용하여 ASVSpoof2019 대회 LA 데이터로 연구된 여러 단일 음성 모형들의 임베딩 벡터들을 가져오고, 이를 합쳐서 새로운 피쳐로 정의한 후, 해당 피쳐에 딥러닝 네트워크를 구축하여 앙상블 모형을 만들어 내는 방식이다.
- 연구되어진 여러 타 단일 모형의 성능들과 비교해 높은 성능을 보였고, 기존에 널리 사용되는 앙상블 방법과의 성능 비교에서도 우위를 점했다.

Figure 1. Proposed Ensemble Architecture



- 각 모델은 사전학습 된 모델의 가중치를 가지고 있다.
- 각 모델의 임베딩은 node의 개수가 class 개수와 맞춰지기 전 마지막 layer의 출력값이다.
- 샘플 하나에 대해 각 모델에서 임베딩 벡터를 추출한다.
- 각 모델의 임베딩 벡터는 길이가 다를 수 있고 임베딩 벡터는 1차원이다.
- 합쳐진 임베딩 벡터를 딥러닝 네트워크로 학습 시킨 후 결과를 얻는다.

Pretrained Model Architecture

1. LCNN Model

- 실험에 사용한 LCNN(Light-CNN)은 Light CNN-9 모델에 몇 개의 layer를 더 추가한 더 깊은 모델이다.
- 32, 48, 64, 32, 32, 32 Convolution channel size와 32, 48, 64, 64, 32 NIN (Network-in-Network) layer을 사용하여 6개의 Convolution layer와 5개의 NIN layer를 반복하는 모델로 pretrained model을 만들었다.
- Light CNN-9 모델에서와 같이 첫 번째 Convolution layer의 kernel size는 5로, 나머지 Convolution layer는 3으로 설정하였다. 첫 번째, 세 번째 Convolution layer를 제외하고는 Batch Normalization을 적용하고 모든 NIN 레이어 뒤에는 Batch Normalization이 온다. Light CNN-9 모델에 정의된 Fully Connected layer를 사용하는 대신 0.5의 확률로 Global Average Pooling layer, Batch Normalization 및 Dropout layer을 사용했다.
- 마지막 Linear layer 전 길이는 32의 임베딩 벡터를 사용한다.

2. OFD Model

- OFD(Overlapped Frequency Distributed) Model은 2개의 스트림이 있는 6개의 블록으로 구성된다.
- 첫 번째 스트림은 주파수와 관련된 feature를 학습하는 것이다. 주파수 축을 따라 n개의 X구역으로 나누고 X구역과 중첩되게 n-1개의 Y구역으로 나눈다. 나눈 구역들을 kernel size 1의 Convolution layer, Batch Normalization, ReLU, 다시 kernel size 1 1의 Convolution layer, Batch Normalization을 지나는 함수 f를 거친 후 element-wise maximum operation으로 겹치는 f(x), f(Y) 중 높은 요소를 취한다.
- 두 번째 스트림은 시간적 feature를 학습한다. Input은 주파수 축을 따라 평균내어져 Channel by Time 크기의 feature map을 만들고 dilation 4의 depthwise Convolution layer를 거친 후 Batch Normalization, Swish activation, point-wise Convolution, ReLU, spatial Dropout을 지나 시간에 대한 feature를 학습하게 된다.
- 마지막 Linear layer 전 임베딩 벡터를 사용하는데, 길이는 128이다.

3. ResMax Model

- RexMax (Residual Network with Max Feature Map) 모델은 9개의 ResMax 블록과 Global Average Pooling으로 구성된다.
- RexMax 블록은 Convolution layer 두 개에 같은 Input을 넣고 두 출력값을 MFM에 통과시켜 Element-wise maximum operation을 수행한다. 그 다음 RexMax 블록의 parameter로 이 작업을 한 번 더 수행할 지 결정한다. 그리고 Input에 대한 Residual을 더해주고 RexMax의 다른 parameter로 Max Pooling을 적용할 지 결정한 다음 Batch Normalization을 해준다.
- ResMax의 임베딩은 Global Average Pooling 전 마지막 벡터이며 길이는 64이다.

3. AASIST Model

- AASIST (AUDIO ANTI-SPOOFING USING INTEGRATED SPECTRO-TEMPORAL GRAPH ATTENTION NETWORKS) 모델은 RawGAT-ST 모델과 공통적인 부분이 있고 이후 AASIST에서 제안하는 구조를 추가한 모형이다.
- RawGAT-ST 와 공통적인 부분은 RawNet2-based encoder로 고차원 representation을 뽑고 이를 주파수 축, 시간 축을 기준으로 Graph attention network와 Graph Pooling으로 구성된 Graph Module에 각각 넣어 출력 graph Gs, Gt를 얻는다.
- AASIST에서 새롭게 제안하는 구조에서 Gs, Gt를 Graph Combination으로 합쳐주고 HS-GAL(heterogeneous stacking graph attention layer)으로 stack node를 여러 개 만들어준 후, MGO(Max Graph Operation)으로 element-wise maximum 출력값을 뽑는다. 뽑힌 출력값의 node-wise maximum, average와 뽑힌 출력값의 stack node를 concatenate하여 마지막 Linear layer로 최종 출력값을 생성한다.
- AASIST의 임베딩은 마지막 Linear layer 전 concatenate된 벡터이며 길이는 160이다.

Performance

Table 1. Classifier Performance

Number	Classifier	EER(%)
1	AASIST	0.83
2	LCNN	3.14
3	ResMax	2.19
4	OFD	5.60

Table 2. Embedding Ensemble Performance

Number	Ensemble Embedding	EER(%) DNN	
		1 Linear	4 Linear
1	AASIST + LCNN	0.44	0.38
2	AASIST + ResMax	0.39	0.41
3	AASIST + OFD	0.76	0.54
4	AASIST + LCNN + ResMax	0.35	0.33
5	AASIST + LCNN + OFD	0.31	0.26
6	AASIST + ResMax + OFD	0.40	0.34
7	AASIST + LCNN + ResMax + OFD	0.33	0.35

Table 2. Soft Voting Ensemble Performance

Number	Ensemble Embedding	EER(%)
1	AASIST + LCNN	0.71
2	AASIST + ResMax	0.73
3	AASIST + OFD	0.78
4	AASIST + LCNN + ResMax	0.65
5	AASIST + LCNN + OFD	0.69
6	AASIST + ResMax + OFD	0.69
7	AASIST + LCNN + ResMax + OFD	0.60

Summary

- 음성 위조 탐지의 문제에서 Embedding 앙상블의 효과는 뛰어났고 기존의 Soft Voting 앙상블 방법보다 좋은 성능을 거뒀다.
- AASIST, LCNN, OFD의 Embedding을 앙상블 했을 때 효과가 가장 좋았고 DNN을 학습시킬 때에 조금 더 깊은 layer를 쌓은 DNN이 성능이 더 좋다는 걸 확인 할 수 있었다. Embedding 앙상블이 간단한 구조에도 잘 동작함을 보였다.
- Embedding 앙상블의 기초적인 모형을 제시했다는 것이 의미가 있으며 구조를 좀 더 고도화 시키는 후속 연구로 발전 시킬 수 있을 것이다.

이 논문은 2022년도 중앙대학교 연구장학기금 지원에 의한 것임.