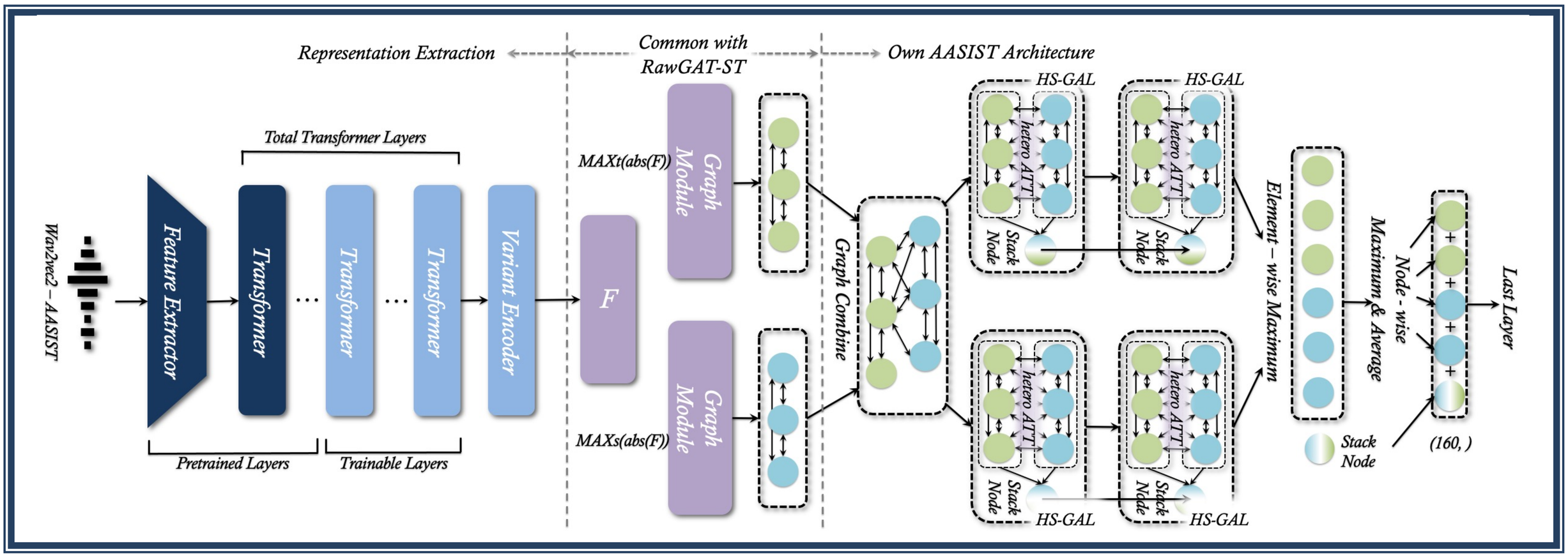
Wav2vec 2.0을 활용한 음성 위조 탐지 모형

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

Figure 1. Proposed Wav2vec2-AASIST Model Architecture



Introduction

- 위조 음성은 실제 음성에 여러 재생산 방법을 사용하여 실제 음성과 유사한 음성을 만들어 낸 것을 의미하며 이를 탐지하여 실제 음성과 구분하는 문제가 음성 위조 탐지이다. 사용자의 음성으로 인증하는 기기가 증가함에 따라 위조 음성과 실제 음성을 구분 하는 모형의 연구가 필요하다.
- 음성위조 탐지모형을 만들기 위해서는 대용량의 음성위조 데이터가 필요한데, 대용량의 사람음성자료로부터 사전학습 된 모형을 이용한다면, 이러한 부분을 보완할 수 있을 것이다. 본 논문에서는 사전 학습 된 Wav2vec 2.0 모델을 활용해 위조 음성을 탐지하기 위한 모형을 제시하고 성능에 대한 결과를 보고한다.
- 모형 성능을 평가하기 위한 척도는 Equal Error rate(EER)을 적용하여 ASVSpoof2019 대회의 LA 데이터에 대해 결과를 냈으며 관련 연구의 타 모형들 성능과의 비교에서 우수함을 보였다.

Model Architecture Explanation

1. Wav2vec2.0 Model

- Wav2vec2(ver. XLS-R 1B)는 약 436,000시간, 128개국 사람의 음성데이터를 사전학습한 모형으로 음성인식, 음성번역 및 화자식별 등 거의 모든 음성 연구 영역에서 탁월한 성과를 보이고 있어 사람 음성 관련 연구들에 많이 활용되고 있다.
- Wav2vec2는 크게 Feature Extractor, Transformer Layer 두 부분으로 나뉘어져 있다. 먼저 Raw audio에서 음소의 단위만큼 Convolution Layer로 특징들을 추출하고 Projection 및 Positional Convolution Embedding을 수행한 뒤, Attention과 Feedforward로 이루어진 여러 개의 Transformer Layer로 구성되어 있다.
- 사전학습 모델의 크기는 Transformer Layer의 개수와 Feedforward의 차원에 따라 결정되며, 실험에 사용한 XLR-R 1B 버전은 Transformer Layer가 총 48개, Feedforward 내의 차원은 1280이다. 제시하는 모델의 실험에서는 Transformer Layer의 개수가 주요한 변수로 작용한다.

2. 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. AASIST는 ASVspoof2019 LA Dataset에서 state-of-the-art 성능을 보였다.
- RawGAT-ST 와 공통적인 부분은 RawNet2에서 사용한 Fixed Sinc Filter와 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 합쳐주어 Heterogeneous Stacking Graph Attention Layer(HS-GAL)으로 stack node를 여러 개 만들어준 후, Max Graph Operation(MGO)으로 element-wise maximum 출력값을 뽑는다. 뽑힌 출력값의 node-wise maximum, average와 뽑힌 출력값의 stack node를 병합하여 마지막 Linear layer로 최종 출력값을 생성한다.

3. Proposed Wav2vec2-AASIST Model

- 우리가 제시하는 모형은 대규모 음성 데이터의 음소에 대해 잘 학습된 사전 모형 Wav2vec2와 기존의 음성 위조 탐지 모형인 AASIST를 결합하여 사전학습 된 음성표현을 위조 탐지 모형으로 새로이 학습하는 구조를 가진다.
- Figure 1에서 볼 수 있듯, 실험의 요소로 사전 학습 된 Wav2vec2에서 Transformer Layer를 얼마나 가져올 것인지, 가져온 Total Transformer Layer에서 몇 번째 Layer까지 Weight를 Freeze 시킬 것인지 변화를 주었다.

- Table 2에서, Total Transformer Layer의 수는 3씩 증가시켜 실험했으며 마찬가지로 Total Transformer Layer 내에서 몇 번 째 Layer까지 Freeze 시킬 건지를 나타내는 Freeze up to N TL도 3씩 증가시켜 실험했다. 한 Total Transformer Layer의 Freeze up to N TL의 결과 중 가장 높은 성능을 Table 2에 정리하였다.

Performance

Table 1. Other System Results On ASVspoof2019 LA Dataset

Number	Classifier	EER(%)
1	AASIST	0.83
2	RawGAT-ST	1.19
3	ResMax	2.19
4	LCNN	3.14
5	RawNet2	3.50
6	OFD	5.60

Table 2. Proposed Wav2vec2-AASIST System Experimental Results

Number	Total TL	Freeze up to N TL	EER(%)
1	3	0	5.93
2	6	3	1.05
3	9	6	0.83
4	12	9	0.20
5	15	3	0.48
6	18	6	0.41
7	21	6	0.46
8	24	3	0.32
9	27	6	0.34
10	30	27	1.92
11	33	3	0.27

Summary

- 제안하는 Wav2vec2 Model과 AASIST Model을 결합시킨 모델의 성능은 음성 위조 탐지 문제인 ASVspoof2019 LA Dataset에서 기존의 Single Classifier들 보다 월등히 뛰어난 결과를 냈다.
- 본래 모델의 Transformer Layer 개수(48)보다 3/4 ~ 1/4 가량 줄였을 때와, Weight Freezing을 0~9 Transformer Layer에 적용시켰을 때에 대체로 좋은 성능을 가져왔다. 이는 대용량의 Pretrained Model이라도 적당한 조절을 통해 사용해야 한다는 것을 알 수 있다.
- 실험의 결과는 128개국의 436,000시간 대용량 음성을 학습한 Wav2vec2 Model이 위조 음성과 실제 사람음성의 구분에도 효과가 있음을 시사하고 잘 학습된 음성 표현을 추출하는 것이 위조 음성 탐지 모델의 고도화에도 중요한 역할을 한다고 볼 수 있다.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science (RS-2023-00208284).