Vol.28, No.4, December 2024
ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)

j.inst.korean.electr.electron.eng.

전기전자학회 논문지
제28권 제4호 2024년 12월
사단법인 한국전기전자학회

[Technical Paper]

[Short Papers]

INSTITUTE OF KOREAN ELECTRICAL AND ELECTRONICS ENGINEERS
http://ikeee.jams.or.kr

Vol.28, No.4, December 2024
ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)

전기전자학회 논문지

j.inst.korean.electr.electron.eng.

제28권 제4호 2024년 12월

[논 문]

[단 편]

사단법인 한국전기전자학회
INSTITUTE OF KOREAN ELECTRICAL AND ELECTRONICS ENGINEERS
http://ikeee.jams.or.kr

Web site : https://ikeee.jams.or.kr
Vol.28, No.4, December 2024

ISSN: 1226-7244 (Print)
ISSN: 2288-243X (Online)

j.inst.Korean.electr.electron.eng.(2024년 12월/제28권 제4호)

목 차

논 문

단 편

* 이 학술지는 정부재원(과학기술진흥기금)으로 한국과학기술단체총연합회의 지원을 받아 출판되었음

ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
j.inst.Korean.electr.electron.eng.Vol.28,No.4,600~606,December 2024
논문번호 24-04-17
https://doi.org/10.7471/ikeee.2024.28.4.600
122

음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN 기반의 Log-Mel 스펙트로그램 이미지 압축 기법

CNN-Based Log-Mel Spectrogram Image Compression Method for Attention Noise Reduction in Speech Emotion Recognition

김 정 윤*, 이 승 호*★

Jeong-Yoon Kim*, Seung-Ho Lee*★

Abstract

This paper proposes convolutional neural networks (CNN) based log-Mel spectrogram image compression method for attention noise reduction in speech emotion recognition (SER) and demonstrates how this method can contribute to improved performance in vision transformer models utilizing attention mechanisms. log-Mel spectrograms, which effectively capture the frequency characteristics of speech signals, are commonly used in SER tasks. In this study, we present a method to reduce attention noise that may arise when processing these spectrograms as images. The core idea is to use a CNN with horizontal kernels to compress the resolution of log-Mel spectrogram images, thereby facilitating the vision transformer model's ability to learn important patterns more effectively. The proposed approach processes the original log-Mel spectrograms at a size of 128×1001 and compresses them into a fixed 128×129 resolution while performing random image interpolation. This preprocessing step aids the model in better extracting relevant features for emotion recognition. This paper propose how the CNN-based compression method preserves essential information from the log-Mel spectrograms while minimizing attention noise in the vision transformer model's attention mechanism. Through experiments using the Crowd Sourced Emotional Multimodal Actors (CREMA) dataset, the proposed method achieved an accuracy of 86.83%, demonstrating superior performance in speech emotion recognition compared to existing methods.

요 약

본 논문은 음성 감정 인식에서 log-Mel 스펙트로그램을 기반으로 한 이미지 압축 기법을 제안하고, 이 기법이 어텐션 메커니즘을 활용한 vision transformer 모델에서 성능 향상에 기여할 수 있음을 보인다. 특히, log-Mel 스펙트로그램은 음성 신호의 주파수 특성을 잘 포착하여 음성 감정 인식에 유용하게 사용되는데, 본 연구에서는 이 스펙트로그램을 이미지 형태로 처리하면서 발생할 수 있는 어텐션 노이즈를 효과적으로 감소시키는 방법을 제시한다. 핵심적인 아이디어는 CNN을 수평 커널로 사용하여 log-Mel 스펙트로그램 이미지의 해상도를 압축하고, 이를 통해 vision transformer 모델에서 중요한 패턴을 보다 효과적으로 학습하도록 돕는 것이다. 제안된 기법은 기존의 log-Mel 스펙트로그램을 128×1001 크기로 처리하고, 이 이미지를 128×129로 고정된 크기로 압축하면서 임의의 이미지 보간이 수행되도록 설계되었다. 이러한 전처리 과정은 모델이 음성 감정 인식에서 유용한 특징을 보다 잘 추출할 수 있도록 돕는다. 본 논문에서는 log-Mel 스펙트로그램의 주어진 특성에 맞게 CNN 기반의 압축 기법을 사용하여 스펙트로그램의 중요 정보를 보존하면서, vision transformer 모델의 어텐션 메커니즘에서 발생할 수 있는 노이즈를 최소화하는 방법을 제안한다. Crowd Sourced Emotional Multimodal Actors(CREMA) 데이터셋을 이용한 실험을 통해, 제안하는 기법이 86.83%의 정확도를 나타내어 기존의 방법들보다 음성 감정 인식에서 더 뛰어난 성능을 보임을 확인하였다.

Key words : Speech Emotion Recognition, Image Compression, Attention Mechanism, Transformer, Deep Learning

* Dept. Electronic Engineering, Hanbat National University
★ Corresponding author
E-mail : shlee@cad.hanbat.ac.kr, Tel : +82-42-821-1137
※ Acknowledgment
This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT)(No. NRF-2022R1F1A1066371)
This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICTChallenge and Advanced Network of HRD) support program(IITP-2024-RS-2022-00156212) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation)
Manuscript received Nov. 25, 2024; revised Dec. 14, 2024; accepted Dec. 18, 2024.

Ⅰ. 서론

음성 감정 인식(Speech Emotion Recognition, SER) [1], [2]은 인간의 감정을 음성 신호를 통해 자동으로 인식하는 기술로, 감정 분석, 인간-컴퓨터 상호작용(HCI), 고객 서비스 시스템, 건강 관리 등 다양한 분야에서 중요한 역할을 하고 있다. 감정은 음성 신호의 억양, 강세, 속도, 톤 등 다양한 특성에 반영되며, 이를 정확하게 추출하고 해석하는 것이 SER의 핵심 과제이다. 최근 SER 분야는 딥러닝 기반의 접근법을 통해 뛰어난 성과를 얻고 있으며, 특히 CNN(Convolutional Neural Networks) [3]과 transformer 모델[4], [5]들이 뛰어난 성능을 보이고 있다. 그러나 음성 감정 인식에서 중요한 도전 과제 중 하나는 음성 데이터의 복잡한 비선형적 특징과 환경적 요인(예: 배경 소음, 화자의 발음 차이 등)에 의해 생성되는 어텐션 노이즈 문제이다. 어텐션 메커니즘은 중요한 정보를 강조하는 데 유용하지만, 과도한 노이즈나 불필요한 정보가 강조될 수 있는 단점이 있다. 이로 인해 감정 인식 성능이 저하될 수 있다. 이러한 문제를 해결하기 위한 다양한 방법들이 제안되고 있지만, 여전히 음성 감정 인식에서의 정확도 향상과 노이즈 감소는 중요한 연구 과제로 남아 있다. 음성 신호의 시간-주파수 정보를 추출하는 방법으로 log-Mel 스펙트로그램은 널리 사용된다. log-Mel 스펙트로그램은 음성의 주파수 특성을 잘 포착할 수 있는 특성 덕분에, 최근 많은 음성 인식 및 감정 인식 시스템에서 주요한 입력 데이터로 활용되고 있다. 특히 Mel-frequency cepstral 계수(MFCC)나 스펙트로그램 기반의 접근법은 감정의 음향적 특징을 잘 반영하는데 유리하다. 그러나 log-Mel 스펙트로그램은 고차원적이고 희소한 데이터로, 이를 효과적으로 처리하기 위한 모델이 요구된다. 최근 연구에서는 이미지 처리 기법을 통해 log-Mel 스펙트로그램을 이미지 형태로 변환하여 시각적 특징을 추출하고, 이를 딥러닝 모델에 입력하는 접근법이 주목받고 있다. 이러한 접근법은 CNN과 같은 모델을 통해 스펙트로그램에서 중요한 시각적 특징을 자동으로 학습할 수 있으며, transformer 계열 모델을 사용한 어텐션 기반의 정교한 분석도 가능하게 한다. 그러나 이미지 압축 및 데이터 해상도의 변화가 모델 성능에 미치는 영향에 대한 연구는 상대적으로 부족하며, 이를 개선하기 위한 방법론이 필요한 상황이다.

따라서 본 논문에서는 음성 감정 인식에서 log-Mel 스펙트로그램을 CNN 기반으로 압축하는 새로운 기법을 제안한다. 제안된 방법은 수평 커널을 이용한 CNN 기반 압축 기법으로, log-Mel 스펙트로그램의 128×1001 크기 이미지를 128×129 고정 크기로 변환하여 임의의 이미지 보간이 자동으로 수행되도록 설계되었다. 이 과정에서 중요한 정보는 보존하면서, 어텐션 메커니즘의 노이즈를 최소화하여, vision transformer 모델[6]을 통한 음성 감정 인식에서 더 나은 성능을 얻을 수 있음을 실험을 통해 입증한다.

Ⅱ. 본론

1. 개요

본 논문에서 제안하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN 기반의 log-Mel 스펙트로그램 이미지 압축 기법의 개요는 그림 1과 같다. 첫 번째 단계로 log-Mel 스펙트로그램 생성 및 전처리를 수행하고 두 번째 단계에서는 CNN 기반 이미지 압축을 수행한다. 마지막으로 vision transformer를 통한 음성 감정 인식을 수행한다.

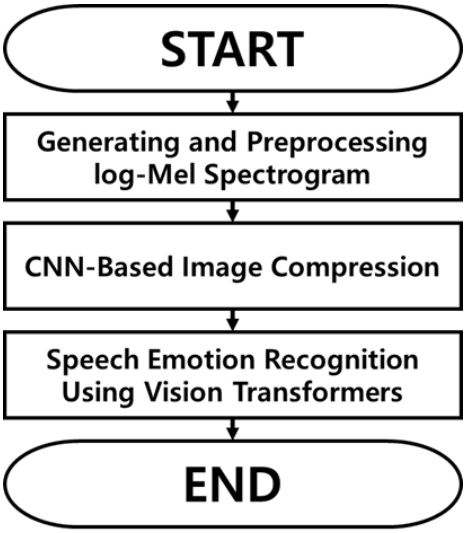

Fig. 1. A overview of the proposed method in this paper.
그림 1. 본 논문에서 제안된 기법의 개요도

2. Log-Mel 스펙트로그램 생성 및 전처리

본 연구에서는 16kHz의 샘플링 레이트를 가진 음성 신호로부터 log-Mel 스펙트로그램을 생성한다. 음성 감정 인식(SER) 작업에서 효과적으로 음성 신호의 시간-주파수 특성을 포착하기 위해, 다음과 같은 파라미터를 사용한다. hop 크기 64, Hamming window function 길이 512, FFT 포인트 수 1024, 그리고 Mel 밴드 수 128. 이러한 설정은 음성 신호의 주파수 해상도를 적절히 반영하며, Mel 스케일이 인간의 청각적 특성을 잘 표현하기 때문에 음성 신호에서 중요한 정보를 잘 추출할 수 있다. 또한, 본 연구에서 사용된 음성 신호의 길이는 대부분 4초 미만이지만, zero padding을 통해 모든 신호의 길이를 4초로 맞춘다. 이를 통해 입력 신호의 길이를 고정시키며, 배치 처리를 위한 일관된 입력 크기를 제공한다. 제로 패딩을 적용한 4초 길이의 각 음성 신호는 앞서 설정한 파라미터에 의해 128×1001 크기의 log-Mel 스펙트로그램으로 변환된다. 여기서 128은 Mel 주파수 밴드의 수를 나타내며, 1001은 시간 축을 따라 계산된 프레임 수이다.

Log-Mel 스펙트로그램을 생성하는 과정은 먼저 short-term Fourier transform(STFT)을 적용하여 음성 신호의 주파수 스펙트럼을 구한다. 그 후 Mel 필터 뱅크를 통해 주파수 축을 Mel 스케일로 변환하고, 마지막으로 로그 변환을 적용하여 Mel 스펙트로그램의 동적 범위를 압축하고, 음성 신호의 중요한 특성을 강조한다. 이 과정에서 사용되는 윈도우 길이 512와 hop 크기 64는 시간 축에서 음성 신호의 중요한 변화를 포착할 수 있는 충분한 해상도를 제공한다. 이와 같은 전처리 과정을 통해 생성된 log-Mel 스펙트로그램은 음성 신호의 중요한 시간-주파수 정보를 잘 보존하며, 이후 모델 학습에 필요한 일관된 입력 크기와 형상을 유지하게 된다. 그림 2는 log-Mel 스펙트로그램 생성 과정을 나타낸다.

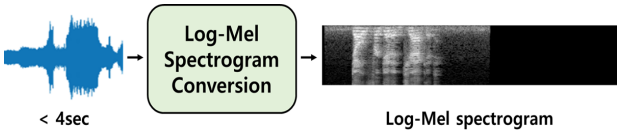

Fig. 2. Process of generating log-Mel spectrograms.
그림 2. Log-Mel 스펙트로그램 생성 과정

본 논문에서 음성 신호를 log-Mel spectrogram으로 변환하는 작업에 사용된 공식은 식 1과 같다.

$$STFT(m,k) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n-aH] \cdot e^{-j\frac{2\pi}{N}kn} \quad (1)$$

주어진 입력 이산 신호 $x[n]$과 길이 L의 윈도우 함수 $w[n]$ (이 경우 Hamming), 홉 크기 H, 그리고 이산 푸리에 변환(DFT) 전체 포인트 수 N이 주어진다. $STFT(m, k)$는 k 번째 frequency bin과 m 번째 시간 프레임에 대한 STFT 계수를 나타낸다. 이후 Mel 스케일로 매핑을 수행하여 log-Mel 스펙트로그램을 생성한다.

3. CNN 기반 이미지 압축

본 연구에서는 log-Mel 스펙트로그램 이미지를 효과적으로 압축하기 위해 CNN 기반의 이미지 압축 기법을 사용한다. 크기가 (i, j) 인 2D 이미지 H와 (k, k) 크기의 2D 커널 F에 대한 CNN의 연산은 식 2와 같다.

$$G[i,j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} H[u,v]F[i-u, j-v] \quad (2)$$

이 방법은 기존의 고차원 log-Mel 스펙트로그램 데이터를 처리하는 데 있어 효율성과 성능을 동시에 고려할 수 있다. CNN을 사용한 이미지 압축의 독창성은 수평 커널을 활용하여 스펙트로그램의 해상도를 축소하는 것이다. 수평 커널을 사용하는 이유는 음성 신호의 시간 축에 대한 중요한 변화를 포착하는 데 유리하기 때문이다. 일반적으로 음성 신호의 감정적 특성은 시간 축에서 발생하는 패턴 변화에 크게 의존하므로, 수평 방향으로의 압축을 통해 중요한 시간적 특성을 유지하면서 해상도를 효율적으로 축소할 수 있다. 이 방식은 이미지의 크기를

128×1001에서 128×129로 줄이며 중요한 정보를 추출하는 데 사용된다. CNN을 통한 압축 과정에서, 각 필터는 원본 스펙트로그램의 중요한 특징을 추출하는 역할을 합니다. 여러 층의 컨볼루션을 통해 생성된 특징 맵은 log-Mel 스펙트로그램의 중요한 시간-주파수 정보를 보존하면서, 불필요한 정보를 제거한다. 각 컨볼루션 층은 고차원적인 데이터를 점차적으로 축소하여, 주요 패턴을 추출하고 압축된 이미지에서 중요한 특성만을 강조한다. 이 과정에서 정보 손실을 최소화하고, 모델의 학습 효율성을 높일 수 있다. CNN을 사용한 이미지 압축의 또 다른 중요한 목적은 어텐션 메커니즘에서 발생할 수 있는 노이즈를 줄이는 것이다. 원본 log-Mel 스펙트로그램은 고차원적이고 희소한 특성을 가질 수 있기 때문에, 어텐션 메커니즘을 사용하기 전 이 데이터를 압축하는 것은 모델이 중요한 정보를 더 잘 학습할 수 있도록 돕는다. 특히, 압축된 이미지에서 불필요한 노이즈를 줄이고, vision transformer 모델이 더 중요한 시간-주파수 특징에 집중할 수 있도록 한다. CNN을 사용한 이미지 압축 기법의 가장 큰 장점은 효율적인 정보 축소와 계산 리소스 절감이다. 고차원 데이터를 그대로 사용하기보다는, CNN을 통해 중요한 특징을 추출하고 압축하여, 모델이 학습하는 데 필요한 계산량을 줄일 수 있다. 또한, CNN은 파라미터 공유와 지역적 연결을 통해 모델의 파라미터 수를 줄이고, 과적합을 방지하는 데 유리하다. 그림 3은 CNN 기반 이미지 압축 과정을 나타낸다. 수평으로 넓은 3×5 크기의 컨볼루션 커널과 시간 축에 대한 stride 2로 이미지를 1차적으로 압축한다. 이후 3×7 크기의 컨볼루션 커널과 시간 축에 대해 stride 3을 적용하여 시간에 대한 주파수 정보를 효과적으로 압축한다. 이후 stride 1인 3×5 컨볼루션 커널을 통해 이미지를 깊게 분석한다. 입력 log-Mel 스펙트로그램의 크기는 128×1001이고 압축된 이미지의 크기는 128×129이다.

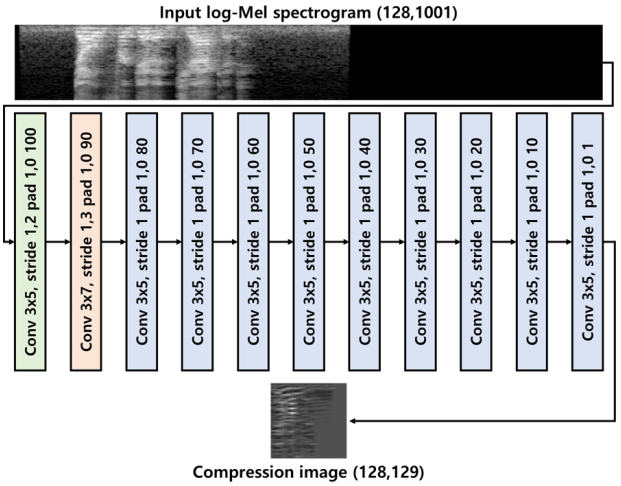

Fig. 3. Process of CNN-based image compression.
그림 3. CNN 기반 이미지 압축 과정

4. Vision transformer를 통한 음성 감정 인식

본 논문에서는 vision transformer를 활용하여 음성 감정 인식(SER) 모델을 구축한다. Vision transformer는 최근 컴퓨터 비전 분야에서 큰 주목을 받으며, 이미지 분류 및 패턴 인식에서 뛰어난 성능을 보여온 모델이다. 전통적인 CNN 기반 모델과 달리, vision transformer는 self-attention 메커니즘을 활용하여 입력 데이터를 처리하며, 이를 통해 공간적, 시간적 관계를 효과적으로 학습할 수 있다. 본 논문에서는 log-Mel 스펙트로그램을 이미지 형태로 변환하고, 이를 vision transformer에 입력하여 음성 신호에서 감정적 특징을 추출하고 인식하는 방식을 채택하였다. 본 연구에서 제시된 CNN 기반의 이미지 압축 기법을 통해 생성된 128×129 크기의 log-Mel 스펙트로그램 이미지는 vision transformer 모델에 입력된다. Vision transformer는 이미지나 시계열 데이터를 작은 패치(patch)로 나누어 각 패치의 정보를 토대로 전체적인 특징을 학습하는 방식으로 동작한다. Log-Mel 스펙트로그램은 주파수와 시간 축에 따라 나누어진 패치들이 모델에 입력될 수 있도록 변환되며, 이를 통해 음성 신호의 감정적 변화를 모델이 잘 인식할 수 있게 된다. Vision transformer의 핵심은 self-attention 메커니즘이다. 이 메커니즘은 입력된 정보에서 중요한 부분을 강조하고, 덜 중요한 부분은 상대적으로 무시하는 방식으로 작동한다. Log-Mel 스펙트로그램의 각 패치는 모델 내에서 attention 값을 계산하여 중요한 주파수와 시간 영역에 집중하게 된다. 이를 통해 모델은 음성 신호의 감정적 특징을 효율적으로 추출할 수 있다. 음성 감정 인식에서는 톤, 억양, 속도 등 시간적 특징과 특정 주파수 대역에서 발생하는 변화를 잘 포착하는 것이 중요하다. Vision transformer는 이처럼 중요한 시간-주파수 영역을 동적으로 선택하여 감정을 인식하는 데 필요한 패턴을 학습한다. ViT는 self-attention 메커니즘을 활용하여 장기 의존성을 학습하는 데 매우 효과적이다. 음성 신호는 짧은 시간 간격 안에서도 감정적인 변화를 나타낼 수 있기 때문에, transformer 모델은 장기적인 시간적 의존성을 모델링하는 데 유리하다. 음성의 억양 변화나 감정의 급격한 전환은 멀리 떨어져 있는 시점들 간의 상호작용을 통해 발생할 수 있다. Vision transformer는 이러한 장기적 의존성을 자연스럽게 학습하며, 보다 정

교한 감정 인식이 가능하다. 식 3은 Attention 메커니즘을 나타낸다.

$$Attention(Q,K,V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \quad (3)$$

셀프 어텐션 레이어는 입력 시퀀스 X로부터 query Q, key K, value V를 도출하기 위해 사용되는 세 개의 학습 가능한 가중치 행렬로 구성된다. 출력 어텐션 $attention(Q, K, V)$는 입력 시퀀스의 가중치이다. K^T는 K의 전치를 나타내며, d_k 는 K의 차원을 나타낸다. 그림 4는 본 논문에서 감정 분류에 사용한 vision transformer의 구조를 나타낸다. Vision transformer는 6개의 헤드, 5층 깊이의 어텐션으로 이루어진 멀티 헤드 어텐션으로 구성된다. 어텐션 메커니즘을 통해 분석이 끝난 특징은 MLP 헤드를 통해 감정을 나타내는 값으로 출력된다.

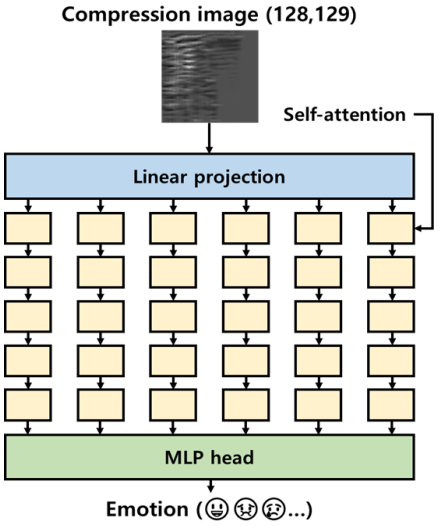

Fig. 4. Process of speech emotion recognition through vision transformer.

그림 4. Vision transformer를 통한 음성 감정 인식 과정

5. 성능 평가

가. 실험 환경

본 논문의 모든 실험은 Windows 워크스테이션의 Windows Subsystem for Linux 2(WSL2) 도커 컨테이너 상에서 수행되었다. Windows 워크스테이션의 사양은 다음과 같다. CPU: Intel(R) Core(TM) i7-10700k 3.8GHz, RAM: 16GB 3200Hz, GPU: NVIDIA GeForce RTX 3070(Video RAM 8GB). 개발도구는 Visual Studio Code와 Pytorch 2.0.1, CUDA 11.8, cuDNN 8.2.1 버전을 사용한다.

Crowd Sourced Emotional Multimodal Actors (CREMA) 데이터 셋 [7]은 7,442개의 다양한 연령, 성별의 배우가 기쁨, 역겨움, 슬픔, 행복, 중립, 놀람 6가지 감정에 대해 연기하는 영상으로 이루어져 있다. 우리는 해당 7,442개의 음성 데이터를 80%:20%의 비율로 학습 셋, 테스트 셋으로 분할하여 각각을 딥러닝 모델의 학습, 딥러닝 모델의 성능평가에 활용한다.

나. 실험 결과

본 논문에서 제안하는 어텐션 노이즈 감소를 위한 CNN 기반의 log-Mel 스펙트로그램 이미지 압축 기법의 객관적인 성능을 평가하기 위해 표 1과 같이 음성 감정 인식 정확도를 비교하였다. 모든 딥러닝 모델은 cross entropy loss를 이용하여 학습되었다. 식 4는 n개의 class에 대한 cross entropy loss의 계산식을 나타낸다.

$$L_{CE} = -\sum_{i}^{n} t_i \log(p_i) \quad (4)$$

식 4의 t_i는 딥러닝 모델의 정답으로 삼는 i에 해당하는 감정 레이블, p_i는 딥러닝 모델 출력의 i 번째 감정에 대한 소프트맥스 확률을 나타낸다.

표 1의 1번 AST [4] 기법과 2번 SepTr [5] 기법은 압축하지 않은 이미지를 이용하여 vision transformer 구조의 네트워크를 통해 음성 감정 인식을 수행한다. 보다 객관적인 성능을 평가하기 위해 인접한 픽셀 값을 이용하는 nearest 보간을 사용하여 128×1001 크기의 이미지를 128×129 크기로 압축한 경우도 평가하여 표 1의 3번에 나타내었다. Nearest 보간으로 이미지 압축한 경우는 입력 이미지가 본 논문에서 사용한 vision transformer로 바로 입력된다. 또한, 이미지 압축을 사용하지 않았을 때의 제안하는 vision transformer의 성능을 표 1의 4번에 나타내었다. 평가 결과 제안하는 CNN 기반의 이미지 압축 기법이 86.83%의 가장 높은 정확도를 나타내었다. CREMA 데이터셋은 거의 동일한 시간에 같은 단어를 발음하는 배우의 음성이 담겨 있는 데이터셋이기 때문에 감정마다 공통된 주파수 패턴이 있을 것으로 추측될 수 있다. 해당 주파수 패턴을 인식하는 작업은 기존의 정사각형 커널을 사용하는 CNN으로도 수행될 수 있다. 그러나 시간 축에 대하여 수평으로 넓은 커널을 사용하는 본 논문의 CNN 기반의 이미지 압축 기법이 log-Mel 스펙트로그램 분석에 더 적합하다고 사료된다. 또한, transformer의 어텐션 메커니즘을 사용한 음성 감정 인식에서 더 높은 정확도를 달성하였다는 것은 감정인식에

중요하지 않은 부분에 대한 잘못된 어텐션으로 인하여 생기는 어텐션 노이즈가 감소되었음을 의미한다. 표 1은 CREMA 데이터셋에 대한 음성 감정 인식 정확도 비교 결과를 나타낸다.

Table 1. Accuracy comparison results about CREMA dataset.

표 1. CREMA 데이터셋에 대한 음성 감정 인식 정확도 비교 결과

No	Method	Accuracy↑
1	AST [4]	67.81%
2	SepTr [5]	70.47%
3	Ours(nearest)	56.41%
4	Ours(no compression)	51.51%
5	Ours(CNN-based)	86.83%

Ⅲ. 결론

본 논문에서는 음성 감정 인식(SER)을 위한 log-Mel 스펙트로그램 이미지 압축 기법을 제안하고, 이를 vision transformer 모델에 적용하여 성능 향상을 달성했다. 제안한 기법은 기존의 SER 모델들과 비교하여 몇 가지 독창적인 특징을 가진다. 첫째, log-Mel 스펙트로그램을 CNN 기반의 수평 커널을 사용하여 압축함으로써, 이미지의 해상도를 효율적으로 축소하고 어텐션 메커니즘에서 발생할 수 있는 노이즈를 줄였다. 둘째, 압축된 이미지는 vision transformer에 입력되어 중요한 패턴을 학습하고, 감정 인식 정확도를 높이는 데 기여한다. 제안된 기법의 성능을 평가하기 위해 CREMA 데이터셋을 사용한 실험을 진행한 결과, 86.83%의 정확도를 나타내어 기존 방법들보다 뛰어난 성능을 보였으며, 유의미한 성능 향상을 확인할 수 있었습니다. 결론적으로 제안된 기법은 음성 감정 인식에서 더 적은 어텐션 노이즈와 더 높은 정확도를 달성하였음을 보였다.

향후 연구에서는 다른 이미지 압축 기법과 어텐션 최적화 방안을 탐구하여, SER 모델의 정확도를 더욱 향상시킬 수 있는 방향으로 더 많은 연구가 필요할 것으로 사료된다.

References

[1] Huang, Zhengwei, et al., "Speech emotion recognition using CNN," *Proceedings of the 22nd ACM international conference on Multimedia.* 2014.

[2] Chen, Lijiang, et al., "Speech emotion recognition: Features and classification models," *Digital signal processing* 22.6 (2012): 1154-1160. DOI: 10.1016/j.dsp.2012.05.007

[3] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.

[4] Gong, Yuan, Yu-An Chung, and James Glass. "Ast: Audio spectrogram transformer." *arXiv preprint* arXiv:2104.01778 (2021).

[5] Ristea, Nicolae Cătălin, Radu Tudor Ionescu, and Fahad Shahbaz Khan. "SepTr: Separable Transformer for Audio Spectrogram Processing." *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.* Vol. 2022. 2022. DOI: 10.48550/arXiv.2203.09581

[6] Dosovitskiy, Alexey, et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale." *International Conference on Learning Representations.* 2020. DOI: 10.48550/arXiv.2010.11929

[7] Cao, Houwei, et al., "Crema-d: Crowd-sourced emotional multimodal actors dataset." *IEEE transactions on affective computing* vol.5, no.4, pp.377-390, 2014. DOI: 10.1109/TAFFC.2014.2336244

BIOGRAPHY

Jeong-Yoon Kim (Member)

2017 : BS degree in Electronic Engineering, Hanbat National University

2019 : MS degree in Electronic Engineering, Hanbat National University

2022~current : Ph. D degree course of Electronic Engineering Hanbat National University

Seung-Ho Lee (Member)

1986 : BS degree in Electronic Engineering, Hanyang University
1989 : MS degree in Electronic Engineering, Hanyang University
1994 : Ph. D degree in Electronic Engineering, Hanyang University
1994~current : Professor, Department of Electronic Engineering, Hanbat National University