논문 2024-4-24 http://dx.doi.org/10.29056/jsav.2024.12.24

딥러닝 모델의 환경음 분류 성능 향상을 위한 시간-주파수 표현

백문기*, 심형섭**†

Time-Frequency Representations for Improving Environmental Sound Classification with Deep Learning Models

Moon-Ki Back*, Hyoung-Seop Shim**†

요 약

스펙트로그램은 오디오 신호처리 연구에서 주파수 성분의 강도를 효과적으로 분석하기 위해 널리 활용하지만, 시간에 따라 변화하는 위상 정보를 표현하는데 한계가 있다. 이러한 한계를 극복하기 위해, 본 논문에서는 파워 스펙트로그램(Power Spectrogram)과 순시 주파수(Instantaneous Frequency)를 결합한 시간-주파수 표현을 탐구하며, 다양한 딥러닝 아키텍처를 활용한 환경음 분류 작업을 통해 효과성을 검증하였다. ESC-50 데이터세트를 대상으로 한 실험에서, 파워 스펙트로그램과 순시 주파수의 수직적 결합 방식을 적용한 ConvNeXt 모델이 87.16%의 분류 정확도를 기록하였으며, 이는 기존 방식 대비 1.7%의 성능향상을 보여준다. 혼동 행렬 분석에서는 물과 관련된 소리와 사이렌 소리가 비슷한 패턴으로 인해 오분류 되었으며, 이를 개선하기 위해 추가 정보가 필요함을 시사한다. 본 연구는 중소규모 데이터세트를 활용하는 오디오 관련 작업에서 딥러닝 모델의 성능을 개선할 가능성을 제시하며, 다양한 소리 관련 응용에 폭넓게 활용될 수 있을 것으로 기대한다.

Abstract

Spectrograms are widely utilized in audio signal processing research to effectively analyze the magnitude of frequency components but they are limited in representing time-varying phase information. To overcome this limitation, this paper explores a time-frequency representation that combines Power Spectrogram (PS) and Instantaneous Frequency (IF) features and validates its effectiveness through environmental sound classification tasks using various deep learning architectures. Experiments on the ESC-50 dataset demonstrate that ConvNeXt model, leveraging the vertical integration of PS and IF, achieves a classification accuracy of 87.16%, reflecting a 1.7% improvement over conventional methods. The confusion matrix analysis reveals that misclassifications often occur for water-related sounds and sirens, as they exhibit highly similar time- frequency patterns, making them challenging to distinguish. This study highlights the potential of the proposed approach to enhance the performance of deep learning models in audio-related tasks, particularly for small- to medium-scale datasets and anticipates broad applicability in sound-related applications.

한글키워드 : 시간-주파수, 환경음, 스펙트로그램, 순시 주파수, 딥러닝

keywords : Time-Frequency, Environmental Sound, Spectrogram, Instantaneous Frequency, Deep Learning

* 한국과학기술정보연구원 선임연구원
** 한국과학기술정보연구원 책임기술원
† 교신저자: 심형섭(email: hsshim@kisti.re.kr)

접수일자: 2024.12.03. 심사완료: 2024.12.12.
게재확정: 2024.12.20.

1. 서 론

소리는 인간과 환경 사이의 중요한 정보 매개체로 작용함으로써, 단순한 감각적 경험을 넘어 과학기술의 발전과 함께 정보전달의 핵심적인 역할로 자리매김했다. 인간은 청각을 통해 눈에 보이지 않는 주변 환경의 변화를 감지하여 위급 상황을 인식할 수 있으며, 다양한 장르의 음악을 통해 언어로 전달하기 어려운 의사소통을 이루어내기도 한다.

소리가 가지는 이러한 특성은 과학기술 발전에 힘입어 다양한 서비스를 등장시켰다. 예를 들면, 핸즈프리(hands-free) 기기에 탑재된 오디오 분류(audio classification) 기술은 집안의 조명이나 온도를 손쉽게 제어함으로써 이동이 제한된 사람들의 삶의 질을 향상시켰고[1], 자동차 관련 장치에 구현된 음성 지원 기능은 쉽고 빠르게 운전자의 요청을 처리하여 편의성을 높였다[2]. 최근에는 인공지능(artificial intelligence)의 발전과 함께 다양한 분야의 오디오 데이터세트가 공개됨으로써, 소리 데이터를 활용한 각종 응용과 연구개발이 크게 늘어나고 있는 추세이다. 대표적으로, HRI(Human-Robot Interaction) 분야에서는 자연스러운 상호작용을 유도하기 위해 소리 데이터를 활용하고 있으며[3], 사고나 재해와 같은 사회적 문제를 저감시키기 위한 목적으로 소리 데이터를 활용하는 연구[4, 5]도 제시되고 있다. 특히, 비언어적(non-verbal) 소리로 분류되는 환경음(environmental sound)은 자연재해, 폭발, 비명과 같은 각종 재난과 위급 상황을 감지하거나, 공장과 설비의 이상을 조기에 발견하여 산업재해를 최소화하는데 기여할 수 있다.

전통적으로 환경음 분류 작업은 소리 데이터가 가지는 고차원적인 특성으로 인하여 기계학습(machine learning) 분야에서 큰 도전 과제로 여겨졌다. 배경 소음이나 비규칙적인 패턴으로부터 유의미한 특징을 추출할 수 있는 정교한 기법이 요구되었으며, 이를 학습할 수 있는 적합한 모델 설계와 학습에 큰 노력이 수반되었다. 그러나 최근 딥러닝(deep learning) 분야의 발전은 이러한 한계를 극복하기 위한 핵심적인 돌파구를 마련했다. 이미지 분류 작업에서 큰 성공을 보인 합성곱 신경망(convolutional neural networks)을 스펙트로그램(spectrogram) 분류로 확장하여, 수작업(handcrafted) 특징 추출보다 합성곱 신경망을 사용하여 소리 데이터를 분석하는 접근이 매우 효과적임이 입증되었다[6]. 게다가, AudioSet[7], ESC-50[8], UrbanSound8K[9]와 같은 대규모 환경음 데이터세트가 공개되면서 오디오 관련 작업에 특화된 여러 딥러닝 모델이 제안되었다.

본 논문은 딥러닝 기반 환경음 분류 작업의 성능향상을 위한 시간-주파수 표현(time-frequency representation)을 탐구한다. 대부분의 관련 연구에서는 서로 다른 소리가 특정 주파수 대역에서 에너지 강도(magnitude)가 다르게 나타난다는 특성에 초점을 맞추어, 모델 입력으로 파워 스펙트로그램(power spectrogram)을 사용한다. 본 논문에서는 시간에 따라 주파수의 위상(phase)이 변화하는 정보를 딥러닝 모델에 제공하기 위하여, 순시 주파수(instantaneous frequency)와 파워 스펙트로그램과 결합하는 새로운 표현을 제안하고, 다양한 딥러닝 모델을 사용하여 제안된 표현이 환경음 분류 성능 개선에 미치는 영향력을 평가한다.

2. 관련 연구

파형(waveform)은 다양한 주파수로 성분으로 구성되어 고유의 음색(timbre)을 가지며, 인간의 청각은 음색의 차이를 바탕으로 서로 다른 소리를 인식한다. 그래서 신호처리 분야에서는 파형으로부터 주파수 성분을 분해하는 푸리에 변환

(Fourier transform)을 사용하여 소리와 같은 각종 신호를 처리하고 분석하는 것이 일반적이다.

그림 1은 소리와 같은 이산-시간 신호(discrete-time signal)를 시간-주파수 표현으로 변환하는 과정을 담고 있다. n 차원 실(real) 벡터 형태의 파형은 이산-시간 푸리에 변환(Discrete-Time Fourier Transform, DTFT)을 통해 복소(complex) 행렬 형태의 스펙트로그램으로 변환될 수 있다. 이를 통해 시간(t)의 흐름에 따라, 주파수의 성분(w) 정보를 확인할 수 있다. 주파수 도메인의 실수부($\Re(X(t, \omega))$)는 주파수에 대한 강도를 나타내고, 허수부($\Im(X(t, \omega))$)는 위상(phase) 정보를 포함하게 된다. 파워 스펙트로그램은 허수부를 제외하여 시간에 따른 주파주 강도를 표현한 것으로, 히트맵(heat map)과 같이 이미지로 시각화하여 직관적인 분석에 널리 사용된다. 딥러닝 모델 관점에서는, 이미지 관련 작업에 널리 활용되는 합성곱 신경망을 통해 파워 스펙트로그램을 분석하기 용이하므로, 다양한 오디오 관련 작업에서 파워 스펙트로그램 입력을 흔하게 찾아볼 수 있다.

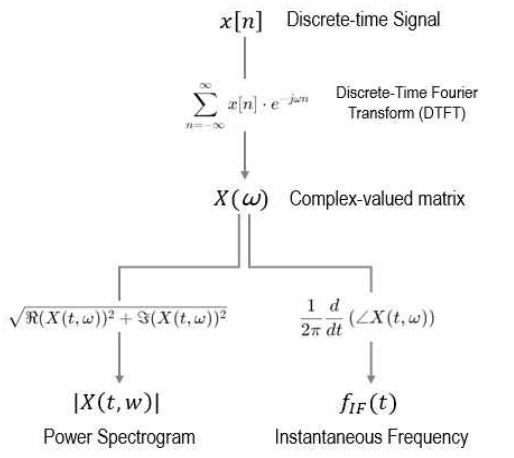

그림 1. 이산-시간 신호를 시간-주파수 표현으로 변환

Fig. 1. Transformation of a discrete-time signal into time-frequency representations

[10]은 합성곱 신경망과 텐서 딥 스태킹 네트워크(Tensor Deep Stacking Network)를 설계하여 ESC-50 데이터세트를 대상으로 환경음 분류를 수행하였다. 각 모델은 스펙트로그램을 입력으로 사용하였으며, 합성곱 신경망이 77%의 분류 정확도를 보이며 더 높은 성능을 보였다.

[11]은 피라미드 결합(pyramidal concatenated) 합성곱 신경망을 제안하였고, 푸리에 변환을 사용해 이미지 형태의 소리 데이터를 입력으로 사용하였다. 피라미드 결합은 스펙트로그램의 차원을 축소하면서 주요 특징을 보존하는 것이 특징으로, 이를 통해 ESC-50 데이터세트에 대하여 81.4%의 분류 정확도를 달성했다.

[12]는 어텐션 메커니즘(attention mechanism)을 합성곱 신경망에 적용하여 환경음을 분류하였다. 특히, 환경음에서 나타나는 하모닉(harmonic) 및 퍼커시브(percussive) 특성을 모델이 학습할 수 있도록 유도하였으며, 그 결과 ESC-50 데이터세트에 대해 84.4%의 분류 정확도를 기록했다.

2020년 Vision Transformer[13]를 통해 자연어 처리에 주로 사용되었던 트랜스포머 계열의 신경망이 비전(vision) 작업에도 높은 효과성이 있음이 밝혀졌다. 이후, 소리 분류 작업에도 이 트랜스포머를 사용해 스펙트로그램을 분석하는 다양한 모델이 제안되었다.

[14]는 다양한 구조의 트랜스포머 블록을 설계하여 실험을 수행하였고, 시간과 주파수를 독립적으로 학습하는 2개의 스트림으로 구성된 블록이 가장 높은 분류 성능을 보인다는 것을 검증하였다(ESC-50 데이터세트에 대하여 57.24%).

[15]는 트랜스포머에 입력되는 스펙트로그램을 패치 수준에서 특징들을 퓨전(fusion)하는 방법을 제안하였고, 이를 통해 ESC-50 데이터세트에 대하여 95.7%의 분류 정확도를 달성했다. 다만, 이 결과는 대규모 데이터세트(AudioSet)를 사전 학습했다는 것을 고려하여 해석할 필요가 있다.

본 논문은 관련 연구의 흐름을 따라 새로운 형태의 시간-주파수 표현을 탐구하고 환경음 분류 작업에서의 효과성을 분석한다.

3. 환경음 분류를 위한 시간-주파수 표현

3.1 환경음 시각화

환경음은 자연적인 환경에서 또는 특정 사건(event)으로 인해 발생하는 넓은 범주의 소리로, 언어적 또는 음악적인 소리와 다른 특성을 가진다. 그림 2는 뇌우, 유리 깨짐, 폭죽, 빗소리에 대한 파워 스펙트로그램과 순시 주파수를 시각화한 것으로, 주파수의 규모뿐만 아니라 연속성에 따라 다른 패턴을 보인다. 즉, 널리 활용되는 파워 스펙트로그램 뿐만 아니라, 순시 주파수 역시 환경음을 정밀 분석하는데 활용될 수 있다. 유리가 깨지는 소리는 짧은 시간 동안 큰 에너지의 소리가 발생하지만 상대적으로 주파수의 연속성은 덜하여 순시 주파수에 표현되는 정보가 매우 한정적이다. 반면, 빗소리의 경우 낮은 주파수 대역에서 일정한 주파수의 소리가 연속적으로 발생하여 시각화된 순시 주파수 상에 선으로서 그 정보가 표현된다.

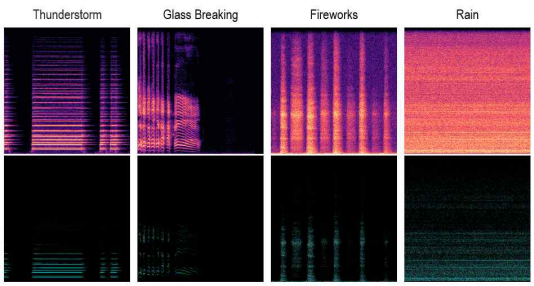

그림 2. 환경음에 따른 파워 스펙트로그램(위)과 순시 주파수(아래) 시각화

Fig. 2. Visualization of the power spectrogram(top) and instantaneous frequency(bottom) for environmental sounds

3.2 딥러닝 모델 학습 방법

그림 2는 파워 스펙트로그램(PS)과 순시 주파수(IF)를 결합하여 하나의 시간-주파수 표현을 생성하고 딥러닝 모델을 통해 환경음을 분류하는 방법을 담고 있다. 구체적으로, 백본 네트워크 F(·)와 다층 퍼셉트론(Multi-Layer Perceptron, MLP) C(·)로 구성된 소리 분류 모델에 그림 1의 시간-주파수 표현을 입력하여 학습을 수행하는 절차를 나타내고 있다.

주어진 파형(n 차원 벡터)에 이산-시간 푸리에 변환을 적용하여 스펙트로그램(n×m 복소 행렬) X를 산출한다. 이후 그림 1의 하단부와 같은 변환 과정을 통해 PS와 IF를 얻는다. 이러한 전처리 과정을 통해 n×m의 실 행렬 형태로 표현된 소리 데이터를 얻을 수 있으며, 행렬의 각 원소(element)를 특정한 범위의 RGB 값으로 매핑함으로써 히트맵 형태로 시각화된 이미지를 생성할 수 있다.

이미지로 표현된 PS와 IF는 크게 3가지 방법으로 결합할 수 있다. 먼저, 두 행렬의 주파수 성분을 나란히 두고 결합하기 위해, 수식 (1)의 HC와 같이 행렬의 열(column)을 기준으로 결합하는 방법이다. 이와 유사하게 두 행렬의 시간 축을 일치시키며 결합하기 위해서는 수식 (2)의 VC와 같이 행렬의 행(row)을 기준으로 결합할 수 있다. 마지막으로, 두 행렬은 이미지와 동일한 포맷이기 때문에 깊이(depth)를 기준으로 결합이 가능하며, 수식(3)의 DC와 같이 α 값을 통해 가중치를 부여하여 결합할 수 있다. 본 논문에서는 α를 0.5로 설정하여 PS와 IF의 가중치를 동일하게 설정하였다.

$$HC = concat_{columns}(PS, IF) = [PS, IF] \quad (1)$$

$$VC = concat_{rows}(PS, IF) = \begin{bmatrix} PS \\ IF \end{bmatrix} \quad (2)$$

$$DC = \alpha \odot PS + (1-\alpha) \odot IF \quad (3)$$

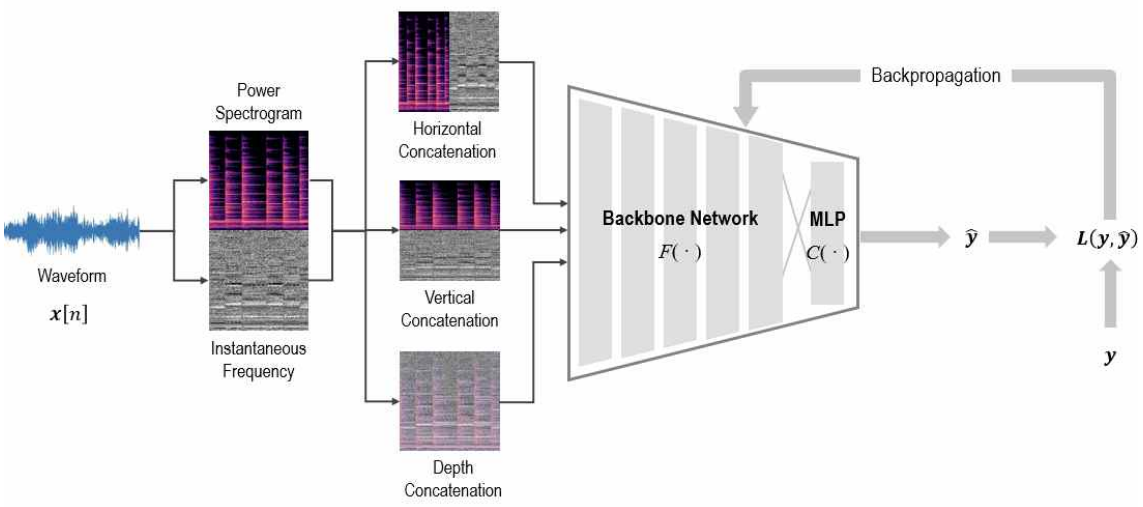

그림 3. 제안하는 시간-주파수 표현 결합 및 딥러닝 모델 학습 절차
Fig. 3. Proposed time-frequency representation concatenation and deep learning model training procedure

DC의 경우, 결합이 완료되어도 행렬의 크기는 n×m으로 유지되나, HC와 VC는 각각 n×2m, 2n×m으로 크기가 변경된다. 일반적으로 딥러닝 모델의 입력은 고정된 크기를 가지므로, HC와 VC는 이중선형 보간법(bilinear interpolation)을 사용하여 모델의 입력과 일치하는 크기로 쉐입(shape)을 변환한다. 본 논문은 3가지 시간-주파수 표현이 딥러닝 모델의 소리 데이터 분류에 미치는 영향력을 분석하는 것이 목표이므로, 추가 매개변수(parameter)를 도입하지 않고 보간법을 활용해 F(·)와 C(·)의 매개변수를 그대로 유지하였다.

$$L(y,\hat{y}) = -\sum_{k}^{K} y^{(k)} \log \hat{y}^{(k)} \qquad (4)$$

앞서 설명한 변환 과정을 통해 본래 파형인 x[n]은 HC, VC, DC의 3가지 시간-주파수 표현으로 변환되었으며, F(·)는 이들을 특징 벡터 z로 매핑하고, C(·)는 z로부터 클래스 레이블 $\hat{y}$을 예측함으로써 소리 데이터 분류가 수행된다. F와 C의 매개변수들은 k개의 레이블로 이루어진 데이터세트 $D=\left\{(x_i,y_i)|_i^N\right\}$에 대하여, 각 데이터 x로부터 예측된 $\hat{y}$과 y를 수식 (4)와 같은 목적 함수(objective function)에 입력하여 최적화가 이루어진다.

3.3 영향력 평가

본 논문에서는 환경음 분류 작업의 대표적인 공개 데이터세트인 ESC-50를 사용하여 제안한 시간-주파수 표현의 영향력을 평가한다. ESC-50은 동물, 자연, 비언어, 일상, 도심 환경에서 수집된 파형들의 집합으로, 총 2,000개의 파형, 50개의 클래스로 구성되어 있다. 모든 파형은 44.1 kHz의 샘플링 레이트(sampling rate)와 5초 길이로 이루어져 있다. 또한, 전체 데이터세트는 5개의 부분 집합(fold)으로 나누어져 있어서, 교차 검증(cross validation)을 통해 평가 결과의 신뢰도를 높일 수 있다.

딥러닝 모델 학습에 앞서, 각 파형은 그림 1의 과정을 통해 PS, IF로 변환하는 전처리 작업을

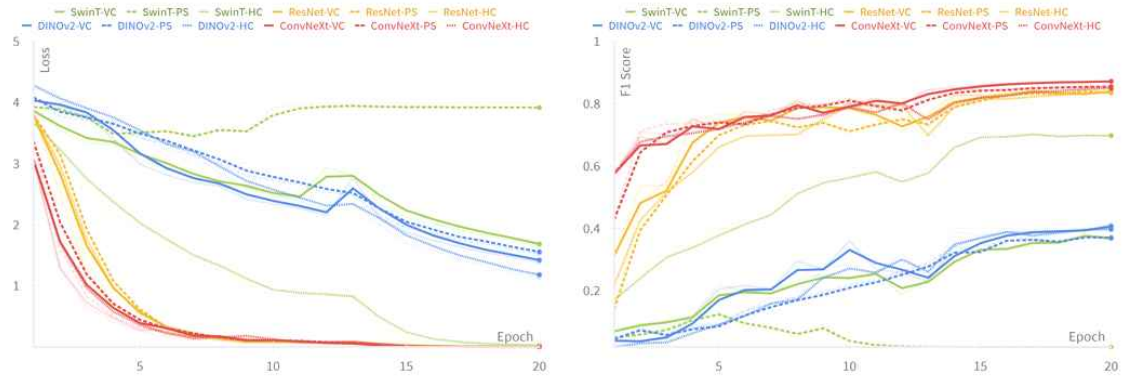

그림 4. 학습 진행에 따른 손실(좌) 및 F1 점수(우) 변화 비교
Fig. 4. Comparison of loss(left) and F1 score(right) changes during training progression

수행하였다. PS는 21.5 ms 단위로, IF는 28.7 ms 단위로 프레임을 이동하며 DTFT를 수행하였고, 이중선형 보간법을 사용해 모두 224×224 크기의 이미지로 변환하여 환경음 분류 모델의 입력으로 사용하였다.

특징 추출을 위한 백본 네트워크 F(·)는 신경망 구조에 따른 차이를 분석하기 위하여, 합성곱 신경망과 Vision Transformer를 모두 고려하였다. 합성곱 계열은 네트워크는 여러 비전 작업에 널리 사용 되어온 ResNet[16]과 비교적 최근 등장한 ConvNeXt[17]를 선정하였고, 트랜스포머 계열의 네트워크 역시 여러 영상 관련 작업에 널리 쓰이는 SWIN Transformer[18]와 방대한 이미지 데이터를 사전 학습하여 기초(foundation) 모델로 자리 잡은 DINOv2[19]를 선정하였다.

ESC-50은 AudioSet과 같은 대규모 데이터세트에 비해 상대적으로 규모가 작기 때문에, 앞서 나열한 모든 모델은 ImageNet[20]을 사용해 사전 학습을 수행하여 시간-주파수 표현의 학습 효율을 높였다. 모든 모델은 AdamW[21]를 사용해 매개변수의 최적화를 진행했고, 학습률(learning rate)은 0.001. 배치 크기(batch size)는 32로 설정하였다. 에포크(epoch)는 20으로 설정하여 학습을 진행하였으며, 모든 성능평가는 5-Fold 교차검증으로 수행하여 5개 결과에 대한 평균을 기록하였다.

그림 4는 각 모델에 3가지 시간-주파수 표현(PS, HC, VC)을 입력하여 학습을 진행하였을 때 나타나는 손실과 F1 점수(F1 score)의 변화를 그래프로 나타낸 것이다. 학습이 진행됨에 따라, 합성곱 계열의 모델은 가파르게 손실을 줄여나가며 최적화되었으나, 트랜스포머 계열의 경우 SWIN Transform에 HC를 입력으로 사용한 경우에만 학습이 이루어졌다. 모든 모델이 ImageNet 사전 학습을 거쳤다는 측면에서, 합성곱 모델이 이미지를 분석하는 작업에 있어 도메인 변화에도 비교적 일반화(generalization) 능력이 우수하다는 것을 확인할 수 있다. 즉, ESC-50과 같이 데이터세트의 규모가 상대적으로 적을수록 합성곱 계열의 모델이 좋은 선택지가 될 수 있다. 트랜스포머 계열 역시 학습이 진행됨에 따라 선형적으로 손실이 낮아지는 것을 확인할 수 있으나, 15 에포크 이후부터는 분류 성능이 정체되는 것을 확인할 수 있다. SWIN Transformer에 HC를 입력으로 사용한 경우, 손실은 합성곱 계열의 모델과 비슷하게 낮아졌으나, F1 점수는 0.69에 그쳤다.

표 1. 시간-주파수 표현에 따른 딥러닝 모델의 환경음 분류 성능 비교
Table 1. Comparison of classification performance of deep learning models for environmental sounds using three time-frequency representations

Backbone Network	Input	Accuracy	Precision	Recall	F1 Score
ResNet # Parameters: 23.5M	PS	0.8350	0.8449	0.8350	0.8399
	HC	0.8360	0.8553	0.8360	0.8455
	VC	0.8390	0.8656	0.8390	0.8521
ConvNeXt # Parameters: 87.5M	PS	0.8547	0.8748	0.8547	0.8646
	HC	0.8491	0.8722	0.8491	0.8605
	VC	0.8716	0.8673	0.8716	0.8694
SWIN Transformer # Parameters: 86.7M	PS	0.0200	0.0004	0.0200	0.0008
	HC	0.6978	0.7014	0.6978	0.6996
	VC	0.3683	0.3587	0.3683	0.3634
DINOv2 # Parameters: 86.5M	PS	0.3704	0.3648	0.3704	0.3676
	HC	0.3969	0.4204	0.3969	0.4083
	VC	0.4066	0.4156	0.4066	0.4110

PS, HC, VC를 네 종류의 백본 네트워크에 입력하여 학습하였을 때 나타나는 평균 성능은 표 1과 같다. 다만, DC의 경우 모든 백본 네트워크에 대해 학습이 이루어지지 못해, 본 영향력 분석에서 제외하였으며, 이는 스펙트로그램을 결합하는 과정에서 주파수의 에너지나 주기성이 손실되어 학습이 불가능한 것으로 해석해 볼 수 있다.

ConvNeXt 백본과 VC를 입력으로 사용했을 때 정확도(accuracy) 87.16%에 F1 점수 0.8694를 달성하였다. PS를 단독으로 학습했을 때 보다 약 1.7%의 정확도 향상이 나타났다. 이러한 결과는 SP만으로는 표현되지 못하는 위상의 변화를 딥러닝 모델이 학습함으로써 나타나는 효과로 해석할 수 있다. 다만, PS와 IF를 어떠한 방식으로 결합하는지에 따라 유의미한 성능 차이가 있음을 주목할 필요가 있다. ResNet과 ConvNeXt 모두 PS와 HC 사이에 유의미한 성능 차이는 발견되지 않았다. 이러한 결과는 커널(kernel) 단위로 이미지 전체를 좌에서 우로 이동하며 특징을 추출하는 합성곱 신경망의 특성에서 비롯된다고 볼 수 있다. VC의 경우, SP 전체에 대한 특징 추출 이후 IF에 대한 특징 추출이 진행되지만, HC의 경우 이질적인 SP와 IF를 번갈아 가며 특징을 추출하기 때문에 IF로 부터 나오는 패턴보다 SP의 패턴에 의존하여 예측을 수행한 것으로 추측해 볼 수 있다.

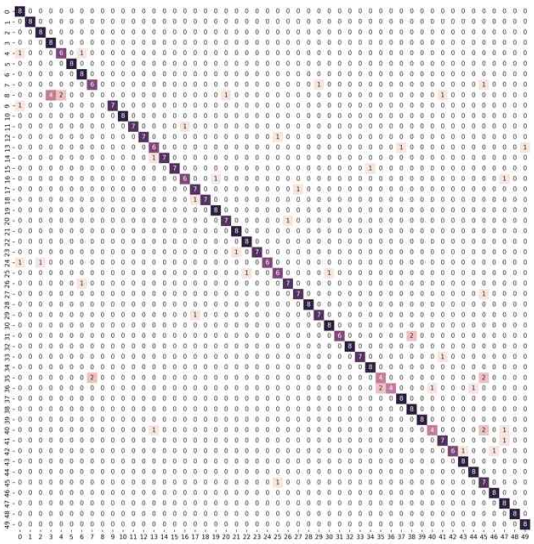

그림 5. 제안된 ConvNeXt-VC의 혼동 행렬
Fig. 5. Confusion matrix of the proposed ConvNeXt-VC

표 2. ESC-50 데이터세트에 대한 본 논문과 선행 연구의 분류 성능 비교
Table 2. Comparison of classification performance on the ESC-50 dataset between prior studies and this work

Method	Accuracy (%)
Human[8]	81.3
CNN[10]	77
TSTF Transformer[14]	57.2
Pyramid-Combined CNN[11]	81.4
TFCNN[12]	84.4
CRNN+Attention[22]	86.5
ConvNeXt-VC(Ours)	87.1

본 논문에서 제안하는 시간-주파수 표현과 선행 연구와의 분류 성능 비교는 표 2와 같다. [14]의 결과 뿐만 아니라 본 논문에서 수행한 트랜스포머 계열의 결과와 같이, 스펙트로그램에 대한 별도의 사전 학습이 없다면 트랜스포머 계열의 모델은 환경음 분류 작업에 적합하지 않다. 반면, [11], [12], [22]와 같이 합성곱 신경망은 전통적인 비전 작업 뿐만 아니라 스펙트로그램 분류 작업도 잘 수행하는 것을 확인할 수 있다. 이러한 결과는 합성곱 신경망이 트랜스포머 구조보다 유도 편향(inductive bias)이 높아서, 이미지 분류에서 스펙트로그램 분류로의 유사 작업 간 도메인 쉬프트(domain shift)가 용이하기에 나타나는 결과로 볼 수 있다.

본 논문은 [7]과 같은 대규모 오디오 데이터세트와 제안하는 시간-주파수 표현을 활용해 합성곱 신경망을 사전 학습하고, 실제 적용 분야(target domain)의 데이터를 사용해 도메인 적응(adaptation)을 수행하는 형태로 본 논문을 활용할 수 있을 것이다. 따라서, 비언어적인 이벤트를 조기에 탐지하기 위한 목적으로, 산업 및 교통 분야의 소음 모니터링, 위험 상황 감지와 같은 서비스가 가능할 것이며, 사람이 밀집된 공간에서 기침 소리와 같은 감염병의 확산 징후를 조기에 감지하는 보건 안전 분야의 서비스로도 확장할 수 있을 것이다. 다만, 그림 5와 같이 아직은 물(water)과 연관된 소리를 잘 분류하지 못하는 한계점이 있다. 물 따르는 소리를 새가 지저귀는 소리와 기차 소리로 잘못 분류한 경우가 많았다. 빗소리의 경우, 물 따르는 소리와 물 내리는 소리로 잘못 분류하였다. 이러한 소리들은 인간의 청각으로 쉽게 구분이 가능한 소리임에도, 주변 환경과 객체 간 상호작용이 더해져 시간-주파수 표현 상에 유사한 패턴이 나타나는 것으로 추측된다. 따라서, 더욱 정밀한 분류 성능을 달성하기 위해서는 이러한 유사 패턴을 구분하기 위한 추가적인 정보를 추출하여 딥러닝 모델에 제공하는 연구가 필요하다.

4. 결 론

본 논문은 환경음 분류 작업을 위한 시간-주파수 표현을 제안하고, 다양한 구조의 딥러닝 모델을 사용해 효과성을 검증하였다. 제안한 시간-주파수 표현은 소리 데이터 분석에 널리 활용되는 파워 스펙트로그램을 기반으로, 소리 데이터의 위상 정보를 담고 있는 순시 주파수를 결합하여 주파수의 에너지와 연속성에 관한 정보를 모두 담고 있다. 합성곱 계열의 네트워크에 제안한 표현을 입력으로 사용하면 환경음 분류 성능을 향상시킬 수 있다. 트랜스포머 계열의 네트워크 역시 제안한 표현을 입력으로 사용하여 유의미한 분류 성능 개선을 도출했으나, 대규모 환경음 데이터세트를 통해 추가적인 실험이 진행될 필요가 있다. 향후 연구에서는 레이블이 없는 소리 데이터에 자기 지도 학습(self-supervised learning)을 적용하여 시간-주파수 표현의 구조와 내재된 패턴을 더욱 잘 이해하는 새로운 모델을 개발하는 방향으로 연구를 확장할 계획이다.

이 논문은 2024년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다. (과제번호: (KISTI)J24JR058-24)

참 고 문 헌

[1] K. Zaman, M. Sah, C. Direkoglu and M. Unoki, “A Survey of Audio Classification Using Deep Learning”, IEEE Access, vol. 11, pp. 106620-106649, Sep. 22, 2023. DOI: 10.1109/ACCESS.2023.3318015

[2] B. Michael, M. Anja, C. Ronee, P. Bastian and A. Florian, “At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces”, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-11, May 02, 2019. DOI: 10.1145/3290605.3300270

[3] B. J. Zhang and N. T. Fitter, “Nonverbal Sound in Human-Robot Interaction: A Systematic Review”, ACM Transactions on Human-Robot Interaction, vol. 12, issue 4, pp. 1-46, Dec. 13, 2023. DOI: 10.1145/3583743

[4] Y. Yuki, C. Premachandra and C. J. Perea, “Audio-Processing-Based Human Detection at Disaster Sites with Unmanned Aerial Vehicle”, IEEE Access, vol. 8, pp. 101398-101405, Jun. 01, 2020. DOI: 10.1109/ACCESS.2020.2998776

[5] G. Ciaburro and G. Iannace, “Improving Smart Cities Safety Using Sound Events Detection Based on Deep Neural Network Algorithms”, MDPI Informatics, vol. 7, no. 3, pp. 23, Jul. 20, 2020. DOI: 10.3390/informatics7030023

[6] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj and T. Virtanen, “DCASE 2017 Challenge setup: Tasks, datasets and baseline system”, DCASE 2017-workshop on detection and classification of acoustic scenes and events, 2017. URL: https://inria.hal.science/hal-01627981/

[7] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal and M. Ritter, “Audio Set: An ontology and human-labeled dataset for audio events”, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Jun. 19, 2017. DOI: 10.1109/ICASSP.2017.7952261

[8] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification”, Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018, Oct. 13, 2015. DOI: 10.1145/2733373.2806390

[9] J. Salamon, C. Jacoby and J. P. Bello, “A Dataset and Taxonomy for Urban Sound Research”, Proceedings of the 22nd ACM international conference on Multimedia, pp. 1041-1044, Nov. 3, 2014. DOI: 10.1145/2647868.2655045

[10] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey and P. Tiwari, “Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network”, IEEE Access, vol. 7, pp. 7717-7727, Jan. 8, 2019. DOI: 10.1109/ACCESS.2018.2888882

[11] F. Demir, M. Turkoglu, M. Aslan and A. Sengur, “A new pyramidal concatenated CNN approach for environmental sound classification”, Applied Acoustics, vol. 170, p. 107520, Jul. 20, 2020. DOI: 10.1016/j.apacoust.2020.107520

[12] W. Mu, B. Yin, X. Huang, J. Xu and Z. Du, “Environmental sound classification using temporal-frequency attention based convolutional neural network”, Scientific Reports, vol. 11, Nov. 3, 2021. DOI: 10.1038/s41598-021-01045-4

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at

Scale", arXiv preprint, Oct. 22, 2020. DOI: 10.48550/arXiv.2010.11929
[14] Y. Zhang, B. Li, H. Fang and Q. Meng, "Spectrogram Transformers for Audio Classification", 2022 IEEE International Conference on Imaging Systems and Techniques (IST), Jul. 20, 2022. DOI: 10.1109/IST55454.2022.9827729
[15] J. Luo, J. Yang, E. S. Chng and X. Zhong, "Vision Transformer based Audio Classification using Patch-level Feature Fusion", 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Dec. 21, 2022. DOI: 10.23919/APSIPAASC55919.2022.9980194
[16] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016. DOI: 10.1109/CVPR.2016.90
[17] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, "A ConvNet for the 2020s", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976-11986, 2022. DOI: 10.1109/CVPR52688.2022.01167
[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012-10022, 2021. DOI: 10.1109/ICCV48922.2021.00986
[19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, ... and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision", arXiv preprint, Apr. 14, 2023. DOI: 10.48550/arXiv.2304.07193
[20] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database", 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255, Aug. 18, 2009. DOI: 10.1109/CVPR.2009.5206848
[21] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization", arXiv preprint, Nov. Jan. 4, 2019. DOI: 10.48550/arXiv.1711.05101
[22] Z. Zhang, S. Xu, S. Zhang, T. Qiao and S. Cao, "Learning Attentive Representations for Environmental Sound Classification", IEEE Access, vol.7, pp.130327-130339, Sep. 4, 2019. DOI: 10.1109/ACCESS.2019.2939495

저 자 소 개

백문기(Moon-Ki Back)

2013.2 충남대학교 컴퓨터공학과 졸업
2021.2 충남대학교 컴퓨터공학과 박사(석·박사통합)
2021.3-2024.7: 한국전자통신연구원 Post-Doc.
2024.8-현재: 한국과학기술정보연구원 선임연구원
<주관심분야> 표현 학습, 데이터 엔지니어링, 멀티 모달 딥러닝, 데이터 증강, 오디오 분류, 이상 탐지

심형섭(Hyoung-Seop Shim)

1999.8 한신대학교 정보통신학과 졸업
2001.8 동국대학교 정보관리학과 석사
2010.8 동국대학교 정보관리학과 박사
2010.10-2012.10 감사원 감사연구원 연구관
2012.12-현재: 한국과학기술정보연구원 책임기술원
<주관심분야> 데이터 공유 플랫폼, 정보관리, 재난 관리 시스템, 국방데이터 관리