한국음향학회지 제39권 제3호 pp. 143~149 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.3 (2020)
https://doi.org/10.7776/ASK.2020.39.3.143

pISSN : 1225-4428
eISSN : 2287-3775

오디오 전처리 방법에 따른 콘벌루션 신경망의 환경음 분류 성능 비교

Comparison of environmental sound classification performance of convolutional neural networks according to audio preprocessing methods

오원근[1†]

(Wongeun Oh[1†])

[1]순천대학교 멀티미디어공학전공

(Received February 25, 2020; revised April 16, 2020; accepted April 22, 2020)

초 록: 본 논문에서는 딥러닝(deep learning)을 이용하여 환경음 분류 시 전처리 단계에서 사용하는 특징 추출 방법이 콘볼루션 신경망의 분류 성능에 미치는 영향에 대해서 다루었다. 이를 위해 환경음 분류 연구에서 많이 사용되는 UrbanSound8K 데이터셋에서 멜 스펙트로그램(mel spectrogram), 로그 멜 스펙트로그램(log mel spectrogram), Mel Frequency Cepstral Coefficient(MFCC), 그리고 delta MFCC를 추출하고 각각을 3가지 분포로 스케일링하였다. 이 데이터를 이용하여 4 종의 콘볼루션 신경망과 이미지넷에서 좋은 성능을 보였던 VGG16과 MobileNetV2 신경망을 학습시킨 다음 오디오 특징과 스케일링 방법에 따른 인식률을 구하였다. 그 결과 인식률은 스케일링하지 않은 로그 멜 스펙트럼을 사용했을 때 가장 우수한 것으로 나타났다. 도출된 결과를 모든 오디오 인식 문제로 일반화하기는 힘들지만, Urbansound8K의 환경음이 포함된 오디오를 분류할 때는 유용하게 적용될 수 있을 것이다.
핵심용어: 환경음 분류, 콘볼루션 신경망, 오디오 특징 추출, 오디오 전처리

ABSTRACT: This paper presents the effect of the feature extraction methods used in the audio preprocessing on the classification performance of the Convolutional Neural Networks (CNN). We extract mel spectrogram, log mel spectrogram, Mel Frequency Cepstral Coefficient (MFCC), and delta MFCC from the UrbanSound8K dataset, which is widely used in environmental sound classification studies. Then we scale the data to 3 distributions. Using the data, we test four CNNs, VGG16, and MobileNetV2 networks for performance assessment according to the audio features and scaling. The highest recognition rate is achieved when using the unscaled log mel spectrum as the audio features. Although this result is not appropriate for all audio recognition problems but is useful for classifying the environmental sounds included in the Urbansound8K.
Keywords: Environmental sound classification, Convolutional neural networks, Audio feature extraction, Audio preprocessing

PACS numbers: 43.60.Bf, 43.60.Lq

I. 서 론

환경음 분류(environmental sound classification)는 가정이나 거리에서 흔히 들을 수 있는 소리를 자동으로 인식하고 분류하는 기술이다. 이 분야에서는 최근 사물인터넷, 원격감시, 홈오토메이션, 또는 청각 장애인이나 노년층을 위한 보조 기구 등의 수요가 높아짐에 따라 인간의 청각과 유사한 수준으로

†**Corresponding author:** Wongeun Oh (owg@scnu.ac.kr)
Department of Multimedia Engineering, Sunchon National University, 255 Jungang-ro, Suncheon-si, Jeonnam 57922, Republic of Korea
(Tel: 82-61-750-3832, Fax: 82-61-750-3830)

인식률을 높이려는 연구가 다양하게 진행되고 있다. 이러한 인식률 향상 방법의 하나로 최근에는 Convolutional Neural Networks(CNN)과 같은 딥러닝 기반의 알고리즘을 적용하는 연구가 활발하게 이루어지는 추세이며 환경음 분류,[1-4] 교통 소음 분류[5] 또는 양서류 울음소리 분류[6] 등에 적용된 예가 있다.

이미지 처리를 목적으로 개발된 CNN을 오디오에 적용하기 위해서는 1차원 데이터인 오디오를 CNN 학습에 적합한 입력으로 만드는 전처리 과정이 필요하다. 전처리 과정에서는 주로 Mel Frequency Cepstral Coefficient(MFCC)나 멜 스펙트로그램과 같이 심리음향에 기반을 둔 시간-주파수 도메인의 오디오의 특징 데이터를 추출한다. 그리고 이를 하나의 이미지로 간주하고 CNN을 훈련하는데, 이러한 방식이 효과적으로 오디오에 적용될 수 있음이 여러 연구를 통해 알려져 있다.

이때 전처리 과정에서 추출하는 오디오 특징은 CNN의 성능에 많은 영향을 준다. 그러나 어떤 특징이 최적인지에 대한 연구는 많지 않고,[3,7] 연구자에 따라 다양한 특성이 사용되고 있다. 예를 들어 Piczak은 로그 멜 스펙트로그램,[1] Tokozume와 Harada[2]는 raw data를 사용하였으며, Boddapati *et al*.[3]은 스펙트로그램, MFCC, Cross Recurrence Plot(CRP)의 조합, 그리고 Su *et al*.[4]은 로그 멜 스펙트로그램과 MFCC에 choroma, spectral contrast, 그리고 tonnetz를 조합한 특징을 사용하였다. 그러나 이들 연구에서는 각각 다른 CNN 구조가 사용되었기 때문에 추출하는 특징에 따른 분류 성능을 직접 비교하기 힘들다는 문제가 있다.

본 논문에서는 환경음 분류 시 전처리 과정에서 사용하는 특징 추출 방법과 설정값에 따라 CNN의 분류 성능을 알아보기 위해 동일한 CNN에 대해 서로 다른 오디오 특징을 추출하여 성능을 비교하였다. 사용한 데이터셋은 환경음 분류 연구에서 많이 사용되는 UrbanSound8K[8]이며, 기존 연구에서 공통적으로 가장 많이 사용되는 멜 스펙트로그램, 로그 멜 스펙트로그램, MFCC, 그리고 delta MFCC를 추출한 다음 3 범위로 스케일링(스케일링 없음, 정규분포 스케일링, -1 ~ 1 범위 스케일링)하였다. 이 데이터를 구조가 다른 4 종의 CNN을 사용하여 분류 성능을 비교하고 가장 우수한 전처리 방법의 조합을 도출하였다. 또한, 이 결과를 잘 알려진 CNN모델인 VGG16과 MobileNetV2에 적용하여 전처리 방법에 따른 인식률을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 사용한 데이터셋에 대한 설명과 특징 추출 방법 그리고 파라미터에 대해서 설명한다. 3장에서는 CNN구조와 실험 결과를 기술하고 4장에서 결론을 맺는다.

II. 데이터셋과 특징 추출

2.1 데이터셋

본 논문에서 사용한 UrbanSound8K 데이터셋은 freesound.org에 업로드된 실제 녹음 파일에서 선별한 10종류의 환경음으로 구성되어 있으며, 다수의 환경음 분류 연구에서 사용된 바 있다[1,3,4,7,9].

이 데이터셋에는 최대 4초 길이의 8732개의 음원이 wav 형식으로 저장되어 있으며, 각 음원의 샘플링 율과 양자화 레벨은 업로드된 원본과 동일하기 때문에 음원별로 다양한 값을 갖는다. 음원의 내용은 도시 일상에서 흔히 들을 수 있는 에어컨 소리, 개 짖는 소리, 자동차 경적 등과 같은 10종류로 구성되어 있다. 소리의 종류와 각 음원의 개수는 Table 1과 같다.

2.2 오디오 데이터 특징 추출

CNN이 오디오를 분류하기 위해서는 먼저 각 음원에서 특징을 추출해야 한다. 오디오의 특징 추출 방

Table 1. The sound classes and number of audio clips in the UrbanSound8K dataset.

classID	Sound class	Number of clips
0	air_conditioner	1,000
1	car_horn	429
2	children_playing	1,000
3	dog_bark	1,000
4	drilling	1,000
5	engine_idling	1,000
6	gun_shot	374
7	jackhammer	1,000
8	siren	929
9	street_music	1,000

법은 여러 가지가 있으나 본 논문에서는 음성 및 음향 인식 분야에서 가장 일반적이면서 기존의 딥러닝을 이용한 음향 분류 문제에서 자주 사용되는 다음 4가지 특징을 추출하여 사용하였다.

- 멜 스펙트로그램
- 로그 멜 스펙트로그램
- MFCC
- MFCC & △MFCC

이상의 특징 추출에는 librosa 라이브러리[10]를 이용 하였으며 상세한 과정은 다음과 같다. 먼저 각 음원은 샘플링 주파수와 양자화 레벨이 다르며 모노와 스테레오가 혼재되어 있으므로 모든 음원을 샘플링 주파수 22050 Hz의 모노 데이터로 변환한다. 다음으로 각 음원을 46.4 ms의 윈도우 단위로 50 %씩 중첩하며 총 174개의 프레임을 구성하고, 각 프레임당 128개의 멜 밴드 에너지를 계산하여 128 × 174 크기의 멜 스펙트로그램 데이터를 추출하였다.

로그 멜 스펙트로그램은 Eq. (1)과 같이 멜 스펙트럼의 파워 S에 로그를 취해 데시벨 값으로 변환하여 구한다.

$$S_{dB} = 10\log_{10}\left(\frac{S}{ref}\right). \quad (1)$$

이때 기준값 ref는 1.0, max 그리고 median 3가지를 사용하였다. 추출한 로그 멜 스펙트로그램 데이터 크기는 128 × 174이다.

MFCC는 음성 및 오디오 처리 분야에서 널리 사용되는 특징으로 본 연구에서는 각 프레임 당 40개의 계수를 사용하여 추출하였다. 이때 MFCC 계산 과정에서 사용되는 로그 멜 스펙트럼을 ref값에 따라 달리 구해서 총 3종의 MFCC값을 구하였다. MFCC 데이터의 크기는 40 × 174이다.

△MFCC는 MFCC 계수의 변화량이며, MFCC와 함께 사용했을 때 인식률을 높이는 효과가 있다. 본 논문에서는 MFCC와 △MFCC를 하나의 배열로 결합하여 80 × 174 크기의 데이터를 입력 데이터로 구성하였다. Fig. 1은 ref = median인 경우 데이터셋에 포함된 소리인 dog_bark에서 추출한 특징과 데이터의 크기를 예시한 것이다.

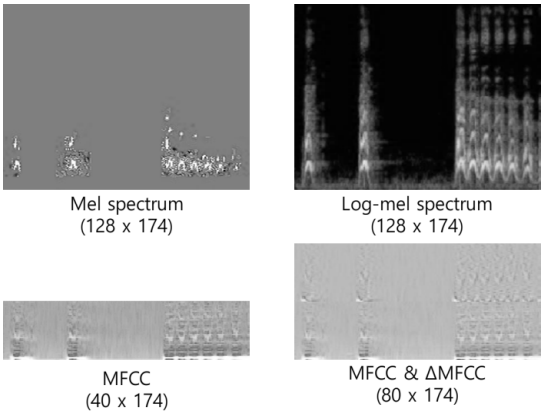

Fig. 1. Extracted audio features of a “dog_bark” sound.

2.3 데이터 스케일링

CNN의 성능은 데이터 개수뿐 아니라 데이터의 분포에도 영향을 받는다. 데이터 분포에 따른 CNN의 성능 변화를 알아보기 위해 4종의 특징 데이터 값 각각을 다음 3가지로 방식으로 스케일링하여 입력 데이터를 구성하고 인식 성능을 평가하였다.

- No scaling : 스케일링 하지 않음
- Standard scaling : 평균 0, 표준편차 1인 정규분포로 스케일링
- Minmax scaling : (-1,1) 범위로 스케일링

III. 실험 및 고찰

3.1 CNN 구조와 학습 방법

전 장에서 추출한 오디오 데이터의 비교를 위해 4개의 CNN을 이용하여 인식률을 실험하였다. 사용한 모델은 사전 실험을 통해 인식률이 비교적 우수한 것을 선별한 것으로서 각각은 레이어의 수, 커널의 수, 풀링(pooling) 그리고 드롭아웃(dropout) 등의 값이 다르게 설정되어있다.

각 모델의 상세한 구조와 파라미터는 Table 2와 같다. 여기에서 Conv2D(k, m, ST = n)는 k개의 필터, m × m 크기의 커널, 그리고 n × n 스트라이드를 사용하는

2D 컨벌루션 레이어를 의미한다. Dense(n)은 n개의 뉴런으로 구성된 완전 연결층(fully connected layer)이며, BN은 Batch normalization, MP(n)은 n × n Max-pooling, 그리고 DO(n)은 비율 n으로 드롭아웃 됨을 나타낸다. 각 뉴런의 활성 함수는 tanh로 표기된 레이어를 제외하고는 relu이며 최종 출력층에는 softmax 함수를 사용하였다.

구현을 위한 코드는 Keras와 Tensorflow를 사용하여 작성하였다. 학습 시 최적화 함수는 Adam 알고리즘[11]을 사용하였으며, 미니 배치(mini batch) 크기는 32, 학습률(learning rate)은 0.002로 최대 300 에포크(epoch)동안 훈련하였다. 또한 매 에포크마다 학습 데이터를 셔플(shuffle)하여 학습하고, 과적합 방지를 위해 검증(validation) 정확도가 10 에포크 동안 개선되지 않으면 학습을 멈추는 조기 종료(early stopping)를 사용하였다. 전체 데이터의 10 %는 학습 시 검증용으로 사용하였으며, 시험 데이터는 10-fold 교차 검증으로 평균 인식률을 구하였다.

Table 2. The architecture of CNN models.

No.	Architecture
CNN1	Conv2D (64, 3, ST = 2), BN, DO (0.4) Conv2D (64, 3, ST = 2), BN, MP (2), DO (0.4) Conv2D (32, 3, ST = 2), BN, DO (0.4) Conv2D (32, 3, ST = 2), BN, DO (0.4) Dense (2048), BN, DO (0.5) Dense (2048), BN, DO (0.5) Dense (10), softmax
CNN2	Conv2D (32, 3, ST = 2), BN Conv2D (32, 3, ST = 2), BN, MP (2), DO (0.5) Conv2D (64, 3, ST = 2), BN Conv2D (64, 3, ST = 2), BN, DO (0.5) Dense (1024), BN, tanh Dense (10), softmax
CNN3	Conv2D (32, 3, ST = 2), BN Conv2D (32, 3, ST = 2), BN, DO (0.5) Conv2D (64, 3, ST = 2), BN Conv2D (64, 3, ST = 2), BN, DO (0.5) Conv2D (128, 3, ST = 2), BN, DO (0.5) Dense (1024), BN, tanh, DO (0.5) Dense (1024), BN, tanh Dense (10), softmax
CNN4	Conv2D (64, 3, ST = 1), BN Conv2D (64, 3, ST = 1), BN, MP (2), DO (0.5) Conv2D (64, 3, ST = 2), BN Conv2D (64, 3, ST = 2), BN, DO (0.5) Dense (1024), BN, DO (0.5) Dense (1024), BN Dense (10), softmax

3.2 MFCC의 ref값에 따른 인식률

Table 3은 Eq.(1)에서 ref 값을 다르게 설정해서 MFCC를 구했을 때 CNN의 인식률이며, 이때 데이터 스케일링은 적용하지 않은 상태이다. ref를 median로 설정하였을 때 ref = 1.0이나 ref = max보다 각각 평균 1.1 %와 2 % 더 나은 인식률을 나타냈다.

이러한 인식률 차이의 원인을 정확히 파악하기는 어려우나 ref 파라미터에 따른 MFCC의 데이터 분포

Table 3. The average 10-fold test accuracy of MFCC according to the 'ref'.

ref \ Model	CNN1	CNN2	CNN3	CNN4	Average
1.0	63.7 %	63.8 %	61.9 %	65.7 %	63.8 %
max	63.7 %	61.9 %	60.9 %	65.0 %	62.9 %
median	64.4 %	65.4 %	63.2 %	66.7 %	64.9 %

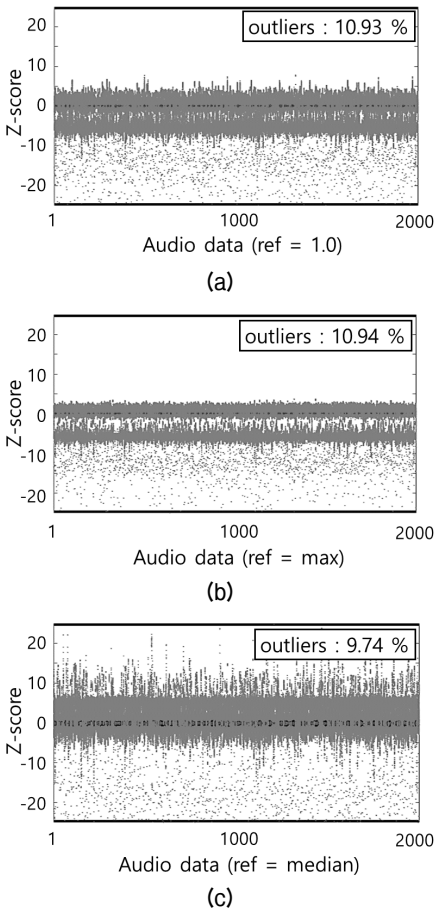

Fig. 2. Z-score distributions of the MFCC data according to the 'ref' parameter (a) ref = 1.0 (b) ref = max (c) ref = median.

가 하나의 요인이 될 수 있을 것이다. Fig. 2는 MFCC 값 분포를 z점수(z-score)로 정규화하여 나타낸 것이다. 그래프의 x축은 8732개의 사운드 데이터에서 임의로 추출한 2000개의 MFCC 데이터이고 y축은 해당되는 z점수를 나타낸 것이다. 그래프에서 ref = 1.0과 max를 사용했을 때 MFCC 데이터는 ref = median보다 더 편중되어 분포하며, 데이터가 비어 있는 silent 구간도 더 많이 나타나는 경향이 있음을 나타내고 있다. 또한 중앙값 절대 편차의 3배 이상 되는 이상점(outlier) 데이터의 비율은 ref = median일 때 9.74 %, ref = 1.0일 때 10.93 %, 그리고 ref = max일 때 10.94 %로 median일 때 이상점의 비율이 다른 것에 비해 1.2 % 정도 낮게 나타났다.

이처럼 ref = median일 때 MFCC의 데이터 분포가 가장 균일하고 이상점이 적게 나타나는 것이 CNN의 인식률에 영향을 주는 하나의 요인으로 추정할 수 있다. 그러나 이것만이 원인이라고 단정하기는 힘들며 정확한 상관관계 분석을 위해서는 데이터 분포와 인식률에 대한 보다 정량적인 연구가 필요할 것으로 생각된다.

3.3 특징별 인식률 비교

Table 4는 오디오 데이터에서 추출한 4가지 특징을 사용해서 구한 인식률을 나타낸 것이다. 이때 스케일링은 적용하지 않은 상태이며, 로그 멜 스펙트럼과 MFCC는 ref = median로 설정하고 추출하였다. 가장 높은 인식률을 보인 것은 로그 멜 스펙트로그램 69 %였으며, 다음으로 68.1 %의 정확도로 MFCC과 delta MFCC를 동시에 사용한 경우의 인식률이 좋은 것으로 나타났다.

Table 4. The average 10-fold test accuracy according to the features.

Features \ Model	CNN1	CNN2	CNN3	CNN4	Average
mel spectrum	58.8 %	24.1 %	23.3 %	27.9 %	33.5 %
log mel spectrum	67.3 %	70.0 %	69.6 %	69.2 %	69.0 %
MFCC	64.4 %	65.4 %	63.2 %	66.7 %	64.9 %
MFCC & △MFCC	67.4 %	68.2 %	66.8 %	69.9 %	68.1 %

3.4 스케일링에 따른 인식률 비교

오디오 특징 데이터의 분포 특성에 따른 CNN의 성능 차이를 실험하기 위해서 추출한 특징을 그대로 사용하는 경우, 평균 0, 표준편차 1인 정규분포로 스케일링한 경우, 그리고 (-1,1) 범위로 스케일링 한 경우에 대해 CNN을 학습하고 인식률을 구했다. Table 5는 각 오디오 특징의 스케일링에 따른 평균 인식률을 나타낸 것이다.

결과를 보면 멜 스펙트럼을 제외한 모든 경우에서 스케일링을 적용하지 않은 경우가 가장 높은 인식률을 나타냈다. 가장 높은 인식률은 로그-멜 스펙트럼을 스케일링하지 않고 사용할 때 69 %이며, 다음으로 MFCC&△MFCC의 인식률이 68.1 %로 높게 나타났다. 이처럼 환경음 인식을 위한 전처리 방식으로 로그 멜 스펙트럼과 MFCC&△MFCC를 스케일링 없이 사용할 때 가장 좋은 결과를 얻을 수 있었다.

특징 중에 멜 스펙트럼은 스케일링 없을 때나 Minmax 스케일링 사용 시의 정확도는 각각 33.5 %와 34.9 %로 매우 낮은 값을 보였다. 그러나 정규분포로 스케일링한 데이터 사용 시에는 62.9 %로 비교적 높은 인식률을 기록했다. 이처럼 멜 스펙트럼은 스케일링과 CNN 구조에 따라 다른 특징에 비해 인식률의 변화가 크게 나타나는 경향을 보였다. 따라서 멜 스펙트럼을 특징으로 사용하여 학습하는 경우에는 정규 분포 스케일링을 우선적으로 고려하고, 또한 다양한 CNN 구조에 대해 인식률을 평가한 후에 적절한 CNN을 선택해서 사용하는 절차가 필요할 것으로 생각된다.

Table 5. The average 10-fold test accuracy of CNN according to the scaling.

Features \ Scaling	No scaling	Standard scaling	Minmax scaling
mel spectrum	33.5 %	62.9 %	34.9 %
log-mel spectrum	69.0 %	66.2 %	64.8 %
MFCC (ref = 1.0)	63.8 %	60.0 %	59.5 %
MFCC (ref = max)	62.9 %	60.7 %	52.8 %
MFCC (ref = median)	64.9 %	60.8 %	58.1 %
MFCC&△MFCC	68.1 %	66.0 %	60.6 %
Average	60.4 %	62.8 %	55.1 %

3.5 이미지넷 CNN을 이용한 분류 성능

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[12]는 이미지넷(ImageNet) 데이터베이스의 영상 인식 성능을 평가하는 대회로서, 2012년 이후에는 딥러닝 네트워크를 이용한 알고리즘이 높은 인식률을 보이며 우승하였다. 앞에서 도출한 오디오 특성을 이미지넷 모델에 적용했을 때도 유사한 결과가 나오는지를 확인하기 위해 2개의 성능이 검증된 이미지넷 CNN 모델을 사용하여 동일한 실험을 수행하였다.

Keras에서 제공되는 이미지넷 모델은 이미지넷에 특화되어 사전에 학습된 것이기 때문에 오디오 데이터에 그대로 사용할 수 없다. 따라서 본 논문에서는 이미지넷 구조에서 출력단을 10개로 바꾼 다음 나머지 전체 네트워크를 다시 학습시켜서 사용하였다. 실험에서 사용한 이미지넷 CNN은 VGG16[13]과 MobileNetV2[14]이다. VGG16은 16층으로 구성된 CNN으로 사용하기 쉬운 구조와 성능을 가지고 있어 많이 사용되는 모델이며, MobileNetV2는 모바일 디바이스와 같은 제한된 환경에서도 사용할 수 있도록 연산량과 네트워크 사이즈를 줄인 모델이다.

입력 데이터인 멜 스펙트럼, 로그 멜 스펙트럼, MFCC 그리고 MFCC&△MFCC 데이터는 VGG16과 MobileNetV2 네트워크의 입력에 맞도록 224 x 224 x 3 크기의 jpg형식의 이미지 데이터로 변환하여 사용하였다. 학습 파라미터로 미니 배치 크기는 32, 학습률은 0.0001, 최적화 함수는 Adam 알고리즘을 사용하고 전체 데이터의 10 %는 검증용으로 사용하여 최대 100 에포크 동안 훈련하였다. 학습 시 과적합 방지를 위해 검증 정확도가 12 에포크 동안 개선되지 않으면 학습을 조기 종료하였다. 또한 인식 성능을 높이기 위해 20 %의 시간축 쉬프트를 적용한 데이터 증강(augmentation)을 사용하였다. 특징 추출은 앞의 실험 결과에서 좋은 결과를 보였던 ref = median과 스케일링하지 않은 데이터를 사용하였다.

실험 결과는 CNN1~CNN4와 유사하게 나타났으며 상세한 결과는 Table 6에 나타내었다. 전반적인 인식율이 앞의 CNN보다 전반적으로 높아진 것은 레이어의 수와 뉴런의 수가 월등히 많은 구조이기 때문이다. 가장 높은 정확도는 로그 멜 스펙트럼을 사용했을 때 얻을 수 있었으며, VGG16은 77.7 %이고 MobileNetV2은 75.6 %의 정확도를 보였다. 두 번째로 높은 정확도는 VGG16에서는 멜 스펙트럼을 사용한 71.6 %이고, MovileNetV2는 MFCC&△MFCC를 사용한 71.9 %로 나타났다.

Table 6. The average 10-fold test accuracy of VGG16 and MobileNetV2.

Features \ Model	VGG16	MobileNetV2
mel spectrum	71.6 %	64.2 %
log mel spectrum	77.7 %	75.6 %
MFCC (ref = median)	69.7 %	69.0 %
MFCC&△MFCC	69.5 %	71.9 %

IV. 결론 및 고찰

본 논문에서는 CNN을 이용한 환경음 데이터 인식 시에 가장 적절한 오디오 전처리 방법에 대해서 실험적으로 고찰하였다. 이를 위해 UrbanSound8K 데이터셋을 이용하여 멜 스펙트럼, 로그 멜 스펙트럼, MFCC, 그리고 MFCC와 delta MFCC를 추출하여 4개의 임의로 구성한 CNN과 2개의 이미지넷에서 검증된 CNN을 사용하여 실험하였다.

그 결과 특징 추출 과정에서 로그 변환 시 ref는 중간값으로 했을 때가 가장 인식률이 좋았으며, 최종 데이터는 스케일링없이 그대로 사용하는 것이 인식률이 높았다. 특징별로는 가장 높은 인식률을 보인 것은 로그 멜 스펙트럼이었으며, 이는 CNN1 ~ CNN4 뿐 아니라 이미지넷 모델에서도 가장 좋은 결과를 보였다. 따라서 환경음 분류 문제를 머신 러닝으로 처리하는 경우에 로그 멜 스펙트럼을 가장 우선적으로 고려할 필요가 있다.

멜 스펙트럼은 CNN1 ~ CNN4에서는 CNN의 구조와 스케일링 방법에 따라 편차가 심했고, VGG16에서는 두 번째로 좋은 성능을 보였으나 MobileNetV2에서는 가장 낮은 성능을 나타냈다. 이와 같이 멜 스펙트럼은 CNN 구조와 파라미터에 따라 편차가 크다고 볼 수 있어서 이를 사용하는 경우 CNN 구조와 스케일링에 대해 충분한 사전 검토가 필요할 것으로 보인다.

이상의 결과는 특정 데이터셋과 6개의 CNN 구조를 이용하여 도출한 것이기 때문에 이를 모든 환경음에 일반화하기에는 한계가 있다. 그러나 본 논문에서 다룬 UrbanSound8K 데이터셋에 포함된 10종의 일상음과 유사한 데이터가 포함된 소리를 인식하는 경우에는 유용하게 적용될 수 있을 것이다.

감사의 글

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2018R1D1A1B07050790).

References

1. K. J. Piczak, “Environmental sound classification with convolutional neural networks,” Proc. IEEE 25th International Workshop on Machine Learning for Signal Processing, 1-6 (2015).
2. Y. Tokozume and T. Harada, “Learning environmental sounds with end-to-end convolutional neural network,” Proc. 2017 IEEE ICASSP. 2721-2725 (2017).
3. V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, “Classifying environmental sounds using image recognition networks,” Procedia Comput. Sci. **112**, 2048-2056 (2017).
4. Y. Su, K. Zhang, J. Wang, and K. Madani, “Environment sound classification using a two-stream CNN based on decision-level fusion,” Sensors, **19**, 1733 (2019).
5. J. Lee, W. Kim, and K. Lee, “Convolutional neural network based traffic sound classification robust to environmental noise” (in Korean), J. Acoust. Soc. Kr. **37**, 469-474 (2018).
6. K. Ko, S. Park, and H. Ko, “Convolutional neural network based amphibian sound classification using covariance and modulogram” (in Korean), J. Acoust. Soc. Kr. **37**, 60-65 (2018).
7. W. Oh, “Audio classification performance of CNN according to audio feature extraction methods” (in Korean), Proc. J. Acoust. Soc. Kr. Supple.2(s) **38**, 64 (2019).
8. J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” Proc. of the 22nd ACM International Conf. on Multimedia, 1041-1044 (2014).
9. J. Salamon and J. P. Bello, “Deep convolutional neural networks and data augmentation for environmental sound classification,” IEEE Signal Process. Lett. **24**, 279-283 (2017).
10. B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, “Librosa: Audio and music signal analysis in python,” Proc. 14th Python Sci. Conf. 18-24 (2015).
11. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980 (2014).
12. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. F. -Fei, “ImageNet large scale visual recognition challenge,” Int. J. Computer Vision, **115**, 211-252 (2015).
13. K. Simonyan and A. Zisseman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556 (2015).
14. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 4510-4520 (2018).

저자 약력

▸ 오 원 근 (Wongeun Oh)

1989년 2월 : 한양대학교 전자통신공학과 학사
1992년 2월 : 한양대학교 전자통신공학과 대학원 석사
1997년 2월 : 한양대학교 전자통신공학과 대학원 박사
1997년 3월 ~ 현재 : 순천대학교 멀티미디어공학전공 교수