JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 5, pp. 913-920, May. 2022

Check for updates

MFCC 기반 환경음 분류 CNN에서 커널 사이즈와 풀링 레이어에 의한 성능분석

정 윤 아[1] · 김 동 회[2*]
[1]강원대학교 IT대학 전기전자공학과 학사과정
[2*]강원대학교 IT대학 전기전자공학과 교수

Performance analysis by kernel size and pooling layer in MFCC-based environmental sound classification CNN

Yun-A Jung[1] · Dong-Hoi Kim[2*]

[1]Undergraduate, Department of Electrical and Engineering, Kangwon National University, Chuncheon, Korea
[2*]Professor, Department of Electrical and Engineering, Kangwon National University, Chuncheon, Korea

[요 약]

오디오 분류에 대한 연구는 청각장애인 및 고령층의 더 나은 삶과 오디오 관련 산업의 발전을 위하여 활발히 진행되고 있다. CNN은 오디오 분류에 사용되는 신경망 중 하나로, 입력 데이터의 특징을 스스로 학습하여 이미지를 분류할 때 주로 활용된다. 커널 사이즈와 풀링은 합성곱 신경망에서 파라미터 개수의 설정에 영향을 미치는 중요한 변수이다. 본 논문에서는 환경음 분류 연구에서 많이 사용되는 오디오 데이터 셋 UrbanSound8K에서 MFCC를 추출하여 CNN에 학습을 진행했다. 세 가지 CNN 시나리오를 설정하고, 각 시나리오에서 커널 사이즈와 풀링 레이어의 개수를 변화시키며 두 값과 정확도 및 파라미터 수의 관계를 알아보는 실험을 진행하였다. 실험을 통해 커널 사이즈와 풀링 레이어의 개수를 증가시킬수록 정확도 및 파라미터가 개선되었음을 확인하였다.

[Abstract]

Research on audio classification is being actively conducted for the improved life of the deaf and elderly, and for the development of audio-related industries. CNN(Convolutional Neural Network) is one of the neural networks used for audio classification, and is mainly used to classify images by learning the characteristics of input data on its own. Kernel size and pooling are important variables that affect the setting of the number of parameters in the CNN. This paper extracts MFCC from UrbanSound8K which is an audio dataset widely used in environmental sound classification studies and makes it learn on CNN. Under three CNN scenarios, we changed the kernel size and the number of pooling layers in each scenario, and tried to find out the relationship between the accuracy and parameter number. Through experiments, it was confirmed that both accuracy and parameters improved as the kernel size and the number of pooling layers increased.

색인어 : 인공지능, 딥러닝, 합성곱 신경망, 파이썬, 소리 분류

Keyword : Artificial Intelligence, Deep Learning, Convolution Neural Network, Python, Sound classification

http://dx.doi.org/10.9728/dcs.2022.23.5.913

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-CommercialLicense(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 06 April 2022; **Revised** 26 April 2022
Accepted 23 May 2022

***Corresponding Author; Dong-Hoi Kim**

Tel: +82-33-250-6349
E-mail: donghk@kangwon.ac.kr

Ⅰ. 서 론

음성 및 오디오의 다중 클래스 분류는 청각장애인과 고령 인구를 대상으로 하는 서비스 및 특정 상황을 감지하는 기술 등 다양한 분야에서 활용되고 있으며 그에 따라 많은 연구가 이루어지고 있다. 소리의 클래스 분류는 인공신경망을 통해 수행되며 이를 통해 청각에서 소외된 계층을 화재, 교통사고 등의 위기 상황으로부터 멀어지게 할 수 있다. 나아가 희귀 동물의 소리를 감지하는 것과 같이 인간에게 한계가 있는 분야의 연구에도 활용되고 있다.

오디오를 분류하기 위해서는 오디오 데이터에서 특징을 추출하여 신경망에 학습시킨 후, 학습 내용을 바탕으로 입력되는 데이터에 대한 클래스 분류를 수행해야 한다. 오디오에서 추출할 수 있는 특징에는 대표적으로 MFCC(Mel Frequency Cepstrum Coefficient)와 Mel spectrogram이 있다. MFCC와 Mel spectrogram은 오디오 데이터의 스펙트럼(Spectrum)으로부터 추출할 수 있다.

오디오 분류는 인공지능의 딥러닝으로 수행될 수 있다. 딥러닝은 패턴 인식 문제 또는 특징점 학습을 위해 다수의 신경층으로 이루어진 모델을 구성하는 기계학습 기술이다. 딥러닝은 더욱 진보된 인공지능 기술로, 인공지능의 발전에 큰 역할을 하였다[1].

합성곱 신경망(CNN; Convolution Neural Network)은 컨볼루션 신경망으로 노드들이 부분적으로만 연결되어 있어서 낮은 복잡도를 가지는 신경망이다. CNN은 컨볼루션 레이어와 풀링 레이어가 반복적으로 배치되어있고 마지막에는 완전연결층을 가지고 있다.

CNN에서는 레이어의 개수가 많아질수록 파라미터 개수와 연산량이 기하급수적으로 증가한다. 이러한 문제는 모바일 디바이스, 임베디드 장치와 같이 저장 공간이 제한되어있는 환경에서 더 심각한 문제가 될 수 있다. 따라서 CNN에서는 학습 파라미터의 개수를 줄이는 것이 중요하다[2][3].

본 논문에서는 오디오에서 추출한 MFCC를 이용하여 CNN을 통해 환경음 분류를 수행하려고 할 때, 커널 사이즈와 풀링 레이어의 개수의 변화에 따른 성능 분석을 하고자 하였으며 실험 결과를 통해 풀링 레이어가 많아지고 커널 사이즈가 증가할수록 정확도가 증가하고 파라미터 개수가 감소하여 정확도와 파라미터 2가지를 모두 만족하는 결과를 얻을 수 있었다.

본 논문의 Ⅱ장에서는 오디오 데이터의 특징과 합성곱 신경망 즉, CNN을 설명하며, Ⅲ장에서는 실험에서 사용된 알고리즘을 소개한다. Ⅳ장에는 실험 환경을 설명하였고 이를 바탕으로 진행한 실험 결과를 Ⅴ장에서 보여준다. 마지막 장에서는 세 가지 시나리오에서 커널 사이즈와 풀링 레이어의 설정에 따라 나타난 실험 결과를 정리한다.

Ⅱ. 관련 연구

2-1 오디오 데이터의 특징

1) waveform

소리는 시계열 데이터로, 음압의 변화를 내포한다. 오디오 데이터는 기본파(fundamental frequency)와 배음(harmonics)로 구성되어있는 여러 주파수가 결합되어있는 형태이기 때문에 오디오 고유의 waveform에서 특징을 추출하기 어렵다. 그림 1은 Urbansound8K 데이터 셋에 포함된 사이렌 소리에 대한 waveform이다.

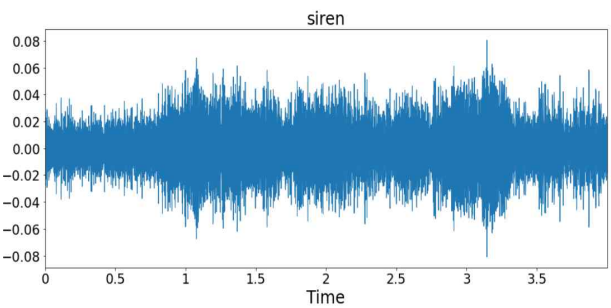

그림 1. 사이렌 소리에 대한 waveform
Fig. 1. Siren waveform

2) Mel spectrogram

사람의 귀는 높은 주파수의 소리에 민감하지 않다. 하지만 소리에서 추출한 고유한 특징을 학습한 신경망을 통해 광범위한 주파수의 소리를 구별해낼 수 있다. MFCC와 Mel spectrogram은 소리의 대표적인 특징으로, 기존의 딥러닝을 이용한 음성인식 및 오디오 분류 문제에서 자주 사용되는 특징에 속한다.

오디오의 waveform에 STFT(Short Time Fourier Transform)를 수행하면 주파수를 x축으로 하는 Spectrum이 생성되고, 이때 y축인 magnitude를 제곱하면 Power spectrum이 생성된다. magnitude에 log scale을 적용하여 데시벨 단위로 변환한 것을 Log spectrum이라고 한다. Log spectrum을 세로로 세워서 프레임마다 쌓으면, 푸리에 변환으로 사라졌던 time domain을 복원할 수 있고, 이를 Spectrogram이라고 한다.

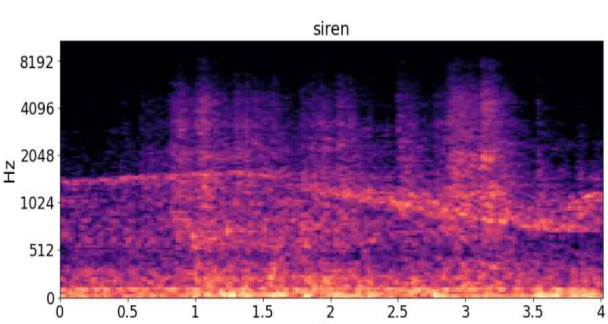

그림 2. 사이렌 소리에서 추출한 멜 스펙트로그램
Fig. 2. Mel-Spectrogram extracted from sirens

Mel filter는 저음의 주파수보다 고음의 주파수에 덜 민감한 사람의 청력에 기반하여 1kHz까지는 선형적으로, 그 이상의 주파수는 log scale로 변환한다. 이와 같은 특성을 가지고 있는 Mel filter를 Spectrogram에 적용시키면 주파수는 Mel frequency로, Power는 log로 맵핑된다. 이와 같이 생성되는 것이 Mel spectrogram이다. 즉, Mel spectrogram은 오디오 신호에 STFT를 가해 얻은 Spectrogram에 Mel filter를 적용하여 얻을 수 있는 특징이다. 그림 2는 사이렌 소리에서 추출한 Mel spectrogram이다.

3) MFCC

MFCC는 Mel spectrogram을 구하는 과정에 이어 log을 취한 뒤 이산 코사인 변환(DCT; Discrete Cosine Transform)을 수행한 것이다. 그리고 주파수가 낮고 정보와 에너지가 몰려있는 12개의 계수(cepstrum coefficient)와 이들을 더한 값을 feature로 사용한다. 즉, MFCC는 Mel spectrogram에 log 및 DCT를 취하고 12개의 계수와 이들로부터 구해진 에너지를 더한 값으로 프레임마다 13개의 값을 feature로 가지는 특징이다. 그림 3은 사이렌 소리로부터 추출한 MFCC이다.

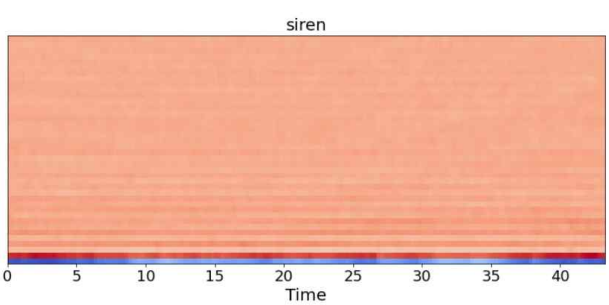

그림 3. 사이렌 소리에서 추출한 MFCC
Fig. 3. MFCC extracted from sirens

2-2 합성곱 신경망 (Convolution Neural Network)

딥러닝의 일종인 합성곱 신경망 즉, CNN은 패턴이나 물체를 인식하는 생물의 시각 처리 과정을 모방한 모형이다. 합성곱 신경망은 LeCun et al.(1998)에 의해 발전하는 계기가 되었다[4].

인공신경망(ANN; Artificial Neural Network)은 딥러닝의 기초가 되는 신경망으로, 인간의 신경망 구조로부터 제작된 알고리즘이다. ANN에는 입력층, 은닉층, 출력층이 있다. 입력층에서 입력된 데이터는 은닉계층에서 활성화 함수를 통해 계산되고, 은닉층의 출력은 가중치 행렬과 활성화 함수로 계산되어 출력층에서 출력된다. CNN과 ANN의 차이점은 CNN에는 컨볼루션 레이어가 있어 합성곱 연산을 바탕으로 특징맵(Feature map)을 추출한다는 것이다. 이 과정을 통해 입력 데이터를 1차원으로 나열하는 ANN과 달리, CNN에서는 입력 이미지의 공간적인 정보를 잃지 않게 된다. 즉, 이미지에서 인접한 픽셀들의 정보를 잃지 않아 입력 데이터의 특징을 파악할 수 있다. 따라서 CNN은 이미지 처리에 적합한 성능을 보인다.

CNN은 사용자의 목적에 맞게 컨볼루션 레이어와 풀링 레이어를 반복적으로 배치시켜 깊은 네트워크를 만들고 마지막에 배치된 완전연결층을 통하여 분류작업을 한다[5]. CNN은 입력층에서 받은 데이터에 합성곱 연산을 수행하여 특징맵을 추출한다. 합성곱 연산은 이미지로부터 특징값을 추출할 때 사용되는데, 주변 화소값에 가중치를 곱한 후, 이들을 더하여 새로운 화소값으로 정한다.

컨볼루션 레이어에서는 입력 이미지의 픽셀 중 커널(kernel)이라는 receptive field 범위 내의 픽셀들을 대상으로 연산을 수행한다. 이미지 전체를 일정한 간격만큼 이동해 가며 커널 사이즈로 정해진 영역에 대하여 계산한다. 따라서, 커널이 작을수록 합성곱 연산 횟수는 증가한다.

풀링 레이어는 정해진 윈도우 사이즈로 stride만큼 이동하며출력값 요약을 통해 데이터의 크기를 줄이는 역할을 한다. 풀링에는 Max Pooling과 Average Pooling이 있다. Max Pooling은 해당 영역의 최댓값으로 요약하고 Average Pooling은 해당 영역의 평균값으로 요약한다. 풀링 레이어를 통해 파라미터 개수를 줄이고 입력 데이터의 잡음이나 왜곡을 해소하는 효과를 얻을 수 있다[1]. 다중 클래스 분류의 경우, 마지막 레이어인 완전연결층은 Softmax 함수를 활성화 함수로 가진다. Softmax 함수는 학습된 항목별로 확률값을 비교하여 분류작업을 할 수 있는 함수이다.

CNN은 주로 이미지 학습에 사용되는 신경망이다. CNN에서는 커널 사이즈가 입력 데이터로부터 특징맵을 제작하는 과정에서 한 번에 어떤 크기를 대상으로 특징을 추출할지를 결정한다. 이미지 분류에서는 커널 사이즈를 작게 설정할수록 좋은 성능을 보이는 연구 결과가 존재한다. 기존 연구에 따르면, 이미지 분류를 수행할 때 커널 사이즈를 작게 설정할수록, 정확도가 개선되는 결과가 도출되었다[6]. 해당 연구는 이미지를 분류하는 CNN에 관한 실험이지만 본 논문에서는 실제 현상을 담은 이미지가 아닌 환경음 즉, 오디오 분류에 대한 실험을 수행하였다. 오디오 분류에서는 실제 현상을 담은 이미지가 아닌, 오디오에서 추출한 특징인 MFCC 혹은 Mel spectrogram을 학습하기 때문에 이미지 분류와 다르게 커널 사이즈를 크게 설정할수록 정확도 성능이 개선된 것으로 나타났다.

제안하는 방법은 환경음으로부터 MFCC를 추출하여 CNN 학습을 통해 분류를 수행하는 알고리즘으로, 실험을 통하여 커널 사이즈를 크게 설정할수록 좋은 성능을 보이는 결과를 확인하였다.

Ⅲ. 제안하는 MFCC 기반 소리 분류 CNN

환경음을 분류하는 방법에는 오디오에서 Mel spectrogram 혹은 MFCC를 추출하여 이를 ANN, CNN 등

의 신경망을 통해 학습하는 방법이 있다. CNN을 사용하여 환경음 분류하는 경우의 오디오 전처리에 관한 기존 연구에서 MFCC를 추출하여 오디오 분류를 수행했을 때 Mel spectrogram을 사용했을 때보다 더 좋은 성능을 보였다[7]. 따라서 본 논문의 실험에서는 오디오에서 MFCC를 추출하였고, 이를 CNN으로 학습하였다. 실험을 진행하여 환경음 분류를 수행할 경우, 커널 사이즈를 크게 설정할수록 성능에 긍정적인 영향을 미치는지 확인하였다. 또한 CNN의 중요한 성능 지표인 파라미터 개수의 결정에 영향을 미치는 요소 중 하나인 풀링 레이어의 개수를 변화시키며 그에 따른 성능변화를 확인하고자 하였다.

아래 그림4는 본 논문에서 진행한 실험의 알고리즘이다. 실험에서는 Urbansound8K 데이터 셋의 오디오 8732개를 불러온 후, 파이썬의 librosa 라이브러리를 이용하여 오디오의 특징으로 MFCC를 추출한다. 추출한 특징을 CNN 신경망에 학습시킨 후, 정확도를 확인한다. 커널 사이즈가 a 미만이거나 풀링 레이어의 개수가 b 미만일 때는 커널 사이즈와 풀링 레이어를 재설정하여 모델을 수정하고 이후 과정을 반복한다. 본 논문의 V에서는 a=6, b=3으로 설정하여 실험을 진행하였다.

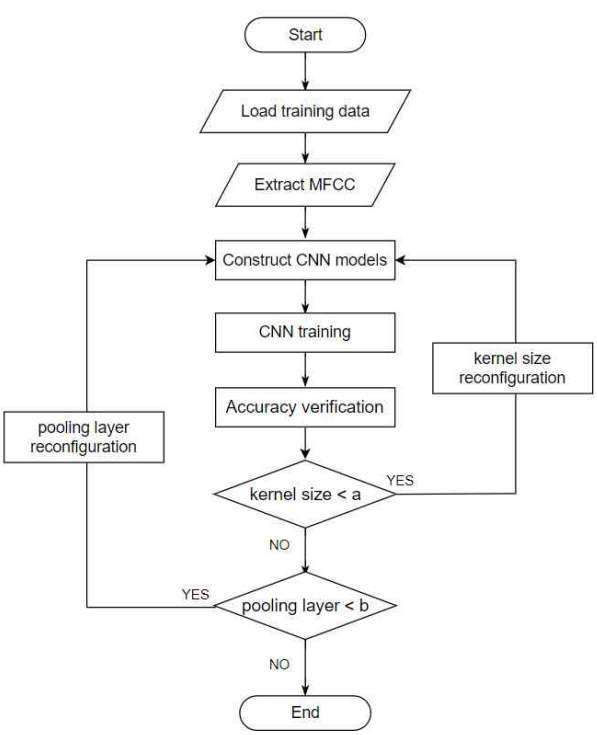

그림 4. 제안하는 알고리즘의 알고리즘
Fig. 4. Flowchart of proposed algorithm

Ⅳ. 실험 환경 및 학습 데이터

4-1 실험 환경

딥러닝을 수행할 수 있는 개발환경에는 주피터 노트북(Jupyter Notebook), 파이참(PyCharm), 비주얼 스튜디오 코드(Visual Studio code), 코랩(Colab; Colaboratory) 등이 있다. 아나콘다는 파이썬의 라이브러리와 주피터 노트북을 제공하여 아나콘다 내부에 설치된 패키지를 웹에서 사용할 수 있도록 한다. 코랩은 클라우드 기반의 주피터 노트북 개발환경으로, 브라우저 내에서 파이썬 스크립트를 작성할 수 있고 GPU를 사용할 수 있다. 본 논문의 실험은 구글에서 제공하는 코랩에서 진행되었다. 아래의 그림 5는 구글 코랩의 구조를 나타낸다.

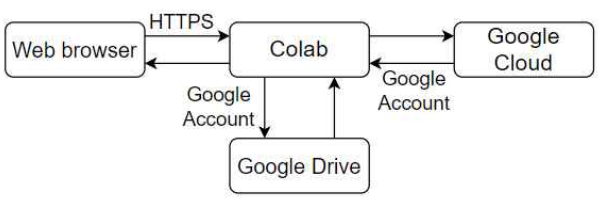

그림 5. 코랩의 구조
Fig. 5. Consturction of Colab

4-2 사용된 데이터 셋

실험에 사용된 데이터는 UrbanSound8K 데이터 셋이다. UrbanSound8K는 NYU에서 수집하고 배포한 오디오 이벤트 자료이다. UrbanSound8K의 데이터는 www.freesound.org로부터 수집된 오디오를 4초 이하로 편집 및 태깅되어있는 자료이다. 이는 10개의 오디오 이벤트 클래스로 구성된 총 8,732개의 오디오 데이터로, 총 9시간 분량의 길이를 가진다[8]. UrbanSound8K 데이터 셋은 환경음 분류 연구에서 다수 사용되고 있다[7][8][9][10]. 해당 데이터 셋을 구성하는 소리의 클래스와 각 클래스를 구성하고 있는 오디오 데이터의 개수는 다음과 같다.

표 1. UrbanSound8K 데이터 셋의 구성
Table 1. UrbanSound8K Dataset Configuration

class	number of data
air_conditioner	1,000
car_horn	429
childern_playing	1,000
dog_bark	1,000
drilling	1,000
engine_idling	1,000
gun_shot	374
jackhammer	1,000
siren	929
street_music	1,000

4-3 MFCC 추출

CNN 신경망에 학습을 진행하기 위해서는 오디오 데이터에서 특징을 추출하는 과정이 필요하다. 기존 연구에서 환경음 데이터의 특징을 MFCC로 추출했을 때의 성능이 Mel spectrogram을 추출했을 때보다 더 나은 수치로 측정되는 것을 확인할 수 있다[8]. 따라서 본 논문의 실험에서는 오디오 데이터 셋에서 MFCC를 추출하여 신경망 학습을 진행하였다. MFCC는 파이썬의 librosa 라이브러리를 이용하여 추출하였다. 실험에서 학습 데이터로부터 추출한 MFCC의 크기는 40×174이다.

4-4 사용된 세 가지 CNN 시나리오

표 2. 세 가지 CNN 시나리오의 컨볼루션 레이어 구성
Table 2. Configuring the Convolution Layer for Three CNN Scenarios

scenario	Description of Convolutional layer
1	Conv,16–Conv,32
2	Conv,16 – Conv,32 – Conv,64
3	Conv,16 – Conv,32 – Conv,64 –Conv,128

본 논문에서는 세 개의 CNN의 구조를 이용하여 실험을 진행하였다. 각각의 신경망으로 진행된 실험을 시나리오 1, 시나리오 2, 시나리오 3이라고 칭했고, 각 시나리오에서 사용된 CNN 구조를 이루는 컨볼루션 레이어의 구성은 표 2와 같다. 표 2에서 Conv는 컨볼루션 레이어를 의미하고, 숫자는 해당 컨볼루션 레이어에서 커널의 개수를 의미한다. 컨볼루션 레이어의 개수는 식 1에 의해 구성되었다. 시나리오 1에서는 두 개 (n=4, 5), 시나리오 2에서는 세 개 (n=4, 5, 6), 시나리오 3에서는 네 개 (n=4, 5, 6, 7)의 컨볼루션 레이어를 사용하였다.

$$2^n (4 \le n \le 7) \tag{1}$$

컨볼루션 레이어의 활성화 함수는 Relu 함수를 이용하였다. 풀링 레이어의 개수는 변수로 두어 실험에서 1개, 2개, 3개로 설정하였고 풀링 레이어의 풀링 사이즈는 2×2로 설정하였다. 완전연결층에서는 Flatten을 통해 1차원 배열로 만든 후, Relu함수를 사용한 Dense와 rate가 0.5인 Dropout 레이어를 각각 두 개 배치하였다. 마지막 Dense 레이어의 활성화 함수는 다중 클래스 분류에 사용되는 Softmax를 사용하였다.

본 논문에서는 세 개의 시나리오에서 그림 4의 알고리즘에 기반하여 컨볼루션 레이어의 커널 사이즈를 2×2부터 6×6까지 증가시키고 각각의 경우에서 풀링 레이어의 개수를 1개, 2개, 3개로 설정했을 때의 오디오 분류 정확도 및 파라미터 수를 관찰하였다.

V. 실험 및 결과

기존 연구에 따르면, 이미지를 분류하기 위한 CNN에서 커널 사이즈를 작게 설정할수록 좋은 성능을 보였다[6]. 이를 바탕으로 본 논문에서는 커널 사이즈 작은 2×2인 경우를 기존 연구 결과로 설정하여 이와 비교하며 실험을 진행하였다.

5-1 시나리오 1의 실험 결과

표 3. 시나리오 1의 정확도 결과
Table 3. Accuracy of scenario 1

filter size \ pooling layer	1	2	3
2×2	41.457	80.882	83.907
3×3	73.692	84.762	86.377
4×4	73.566	85.415	86.100
5×5	80.147	85.213	84.827
6×6	81.740	84.291	85.675

표 2와 같이 컨볼루션 레이어를 구성하고, 풀링 레이어와 커널 사이즈를 증가시키며 실험한 결과, 정확도는 위와 같이 측정되었다. 시나리오 1의 경우, 표 3에서 보는 바와 같이 정확도가 90%를 넘지 못하였다. 하지만 기존 연구 결과로 커널 사이즈가 작은 2×2을 기준으로 비교했을 때, 본 논문에서는 커널 사이즈를 증가시킬수록 정확도는 그에 비례하여 증가하였다. 또한, 같은 커널 사이즈인 경우에도 풀링 레이어의 개수를 늘릴수록 정확도는 증가하였다.

파라미터 개수는 딥러닝의 성능을 평가하는 중요한 지표 중 하나이다. 딥러닝에서는 파라미터의 개수를 줄이는 것이 중요한 고려사항이 된다. CNN의 컨볼루션 레이어가 증가할수록 파라미터 개수가 증가하게 되는데, 이로 인해 용량의 문제점이 발생 될 수 있다. 학습 파라미터의 개수가 많아질수록 학습이 어렵고 메모리가 부담해야 하는 용량도 커진다. 저장 공간의 크기가 제한되어있는 경우, 학습 파라미터의 개수가 많으면 메모리가 감당할 수 있는 용량이 초과되어 신경망 코드와 학습 파라미터를 업로드하지 못하는 상황도 발생할 수 있다. 따라서 CNN에서는 성능은 유지하면서 학습 파라미터의 개수를 줄이는 것이 중요하다[3].

실험에서 컨볼루션의 파라미터의 개수 P는 아래 식 2와 같이 입력 이미지와 커널의 정보에 의하여 구할 수 있다.

$$P = (K^2 \times C \times N) + N \tag{2}$$

식 2에서 K는 컨볼루션 레이어에서 사용된 커널 사이즈이고, C는 입력 이미지의 채널, N은 커널의 개수이다.

시나리오 1의 실험에서 파라미터 개수는 다음과 같다. 커널 사이즈와 풀링 레이어의 개수를 증가시킬수록 파라미터의 개수는 감소하는 것으로 나타났다.

표 4. 시나리오 1의 파라미터 개수
Table 4. Number of parameters in scenario 1

filter size \ pooling layer	1	2	3
2×2	1,569,242	1,557,418	353,194
3×3	1,467,434	1,395,786	363,594
4×4	1,268,378	1,209,642	279,850
5×5	1,174,826	1,199,690	298,570
6×6	990,170	1,034,154	309,162

5-2 시나리오 2의 실험 결과

표 5. 시나리오 2의 정확도 결과
Table 5. Accuracy of scenario 2

filter size \ pooling layer	1	2	3
2×2	80.481	86.789	87.646
3×3	84.049	88.566	89.797
4×4	86.571	89.410	89.783
5×5	85.604	90.212	89.223
6×6	85.874	90.326	–

표 6. 시나리오 2의 파라미터 개수
Table 6. Number of parameters in scenario 2

filter size \ pooling layer	1	2	3
2×2	5,860,218	1,354,618	338,810
3×3	5,062,090	1,007,050	269,770
4×4	3,925,050	664,634	197,690
5×5	3,219,146	507,594	138,954
6×6	2,215,290	232,826	–

시나리오 2에서 정확도와 파라미터 개수는 각각 표 5, 표 6과 같이 측정되었다. 시나리오 2에서 풀링 레이어의 개수를 3개로 설정하였을 때는 차원 문제로 인하여 커널 사이즈가 6x6인 실험은 진행하지 못하고 5×5까지 증가시키며 실험을 진행하였다. 기존 연구 결과에 근거하여 작은 커널 사이즈인 2×2인 경우를 기본 연구대상으로 설정하여 결과를 비교하였다.

시나리오 2에서 커널 사이즈와 풀링 레이어의 개수를 크게 설정할수록 정확도가 대략 90%까지 증가하는 것을 확인하였다. 시나리오 1에 비하여 작은 폭으로 개선되었지만, 작은 커널 사이즈가 2×2일 때의 정확도와 비교하면 정확도가 증가함을 확인하였다. 또한 풀링 레이어의 개수를 1개에서 3개로 늘릴수록 정확도가 증가하는 경우도 확인할 수 있다. 표 6에서 나타난 결과와 같이 커널 사이즈와 풀링 레이어의 개수를 늘릴수록 파라미터 값도 작아지는 결과를 얻을 수 있었다.

아래 그림 6과 그림 7은 시나리오 2에서 풀링 레이어를 2개로 설정했을 때, 정확도가 90% 이상으로 측정된 경우의 정확도 그래프이다. 그림 6과 그림 7은 각각 커널 사이즈를 5×5와 6×6으로 설정했을 때의 결과이다. 가로축은 학습의 진행을 나타내는 에포크 수(Epochs)이고, 세로축은 정확도(Accuracy)이다.

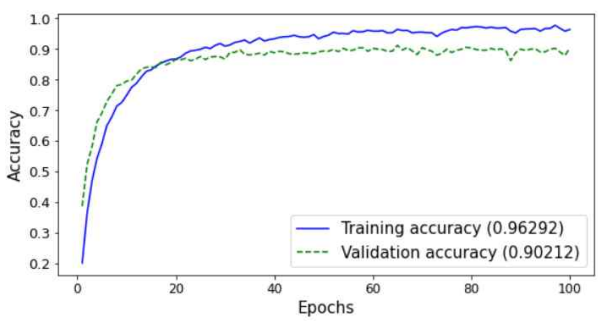

그림 6. 커널 사이즈 5×5의 정확도
Fig. 6. Accuracy of kernel size 5x5

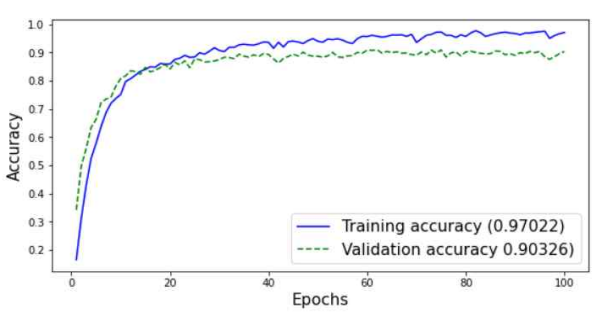

그림 7. 커널 사이즈 6×6의 정확도
Fig. 7. Accuracy of kernel size 6x6

5-3 시나리오 3의 실험 결과

표 7. 시나리오 3의 정확도 결과
Table 7. Accuracy of scenario 3

filter size \ pooling layer	1	2	3
2×2	86.587	88.695	90.441
3×3	87.931	90.326	92.044
4×4	87.708	91.299	–
5×5	87.471	–	–
6×6	88.977	–	–

표 8. 시나리오 3의 파라미터 개수
Table 8. Number of parameters in scenario 3

filter size \ pooling layer	1	2	3
2×2	21,802,682	4,632,250	978,618
3×3	17,137,930	2,588,938	393,482
4×4	11,380,602	747,386	–
5×5	7,447,050	–	–
6×6	2,650,298	–	–

시나리오 3에서의 정확도 결과와 파라미터 개수는 각각 표 7, 표 8과 같다. 다운 샘플링(down sampling)을 하는 과정에서 출력의 사이즈가 0 이하가 되는 문제점이 발생하여 풀링 레이어가 2개일 때, 커널 사이즈가 5×5와 6×6인 경우와 풀링의 개수가 3개일 때, 커널 사이즈가 4×4, 5×5와 6×6인 경우의 실험은 진행하지 못하였다. 따라서 커널 사이즈를 풀링 레이어의 개수가 2일 때는 4×4까지, 풀링 레이어가 3개인 경우에는 3×3까지 증가시키며 실험을 진행하였다.

시나리오 3의 경우, 시나리오 1과 시나리오 2에 비하여 비교적 작은 커널 사이즈부터 정확도가 높게 측정되었지만 가장 큰 파라미터 개수를 가진다. 모든 시나리오에서 종합적으로 살펴보면 커널 사이즈와 풀링 레이어가 증가할수록 정확도는 증가하고 파라미터 개수는 감소한다. 시나리오 1은 작은 파라미터 개수를 가지지만 90% 미만의 낮은 정확도를 가지고 시나리오 3의 경우는 정확도는 높게 측정되었지만, 시나리오 2에 비하여 파라미터 개수가 상당히 큰 값으로 나타나는 것을 확인할 수 있었다. 시나리오 2의 결과는 시나리오 3의 정확도와 비슷하면서 파라미터 개수는 시나리오 3의 결과보다 작았다. 따라서 CNN을 통하여 오디오 분류를 수행할 때, 신경망을 깊게 구성하지 않아도 커널 사이즈 및 풀링 레이어의 개수를 적절히 증가시킨다면 시나리오 2와 같이 90%정도의 정확도와 작은 파라미터 수 두 가지를 모두 만족할 수 있다. 즉, 시나리오 3과 같이 컨볼루션 레이어가 많은 신경망과 성능은 비슷하게 유지하되 파라미터 개수를 줄이고자 한다면, 커널 사이즈와 풀링 레이어의 개수를 적절하게 조정하여 구현할 수 있다. 정확도를 높이기 위해 컨볼루션 레이어를 많이 쌓은 알고리즘이 있다면, 커널 사이즈와 풀링 레이어의 개수의 값을 변화시켜 컨볼루션 레이어의 개수를 줄이면서 성능을 개선시킬 수 있을 것이다.

아래 그림 6과 그림 7은 시나리오 3에서 풀링 레이어가 2개인 경우, 정확도가 90% 이상으로 측정되었을 때의 정확도 그래프로, 각각 커널 사이즈를 3×3과 4×4로 설정했을 때의 결과이다. 그림 10과 그림 11은 풀링 레이어가 3개인 경우 정확도가 90% 이상으로 측정된 경우의 정확도 그래프로, 각각 커널 사이즈를 2×2와 3×3로 설정하였을 때의 결과이다.

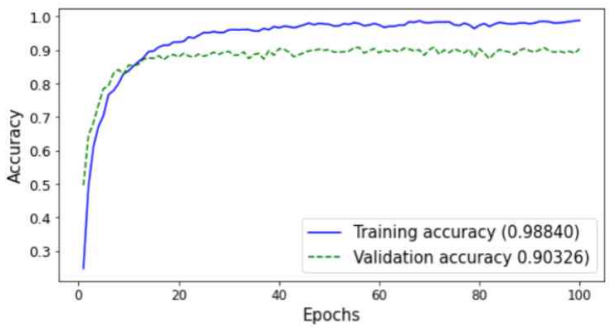

그림 8. 커널 사이즈 3×3의 정확도
Fig. 8. Accuracy of kernel size 3x3

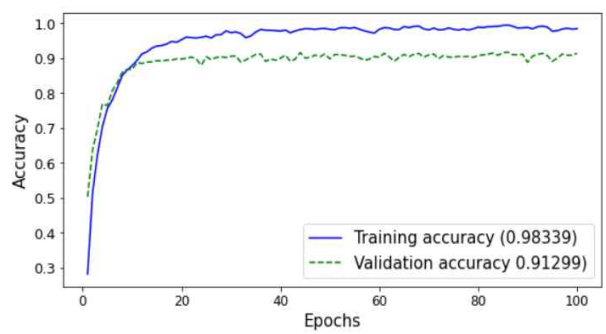

그림 9. 커널 사이즈 4×4의 정확도
Fig. 9. Accuracy of kernel size 4x4

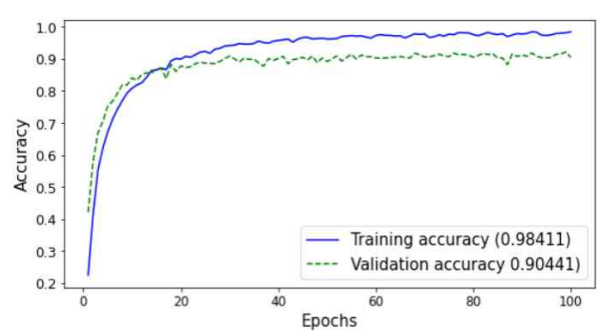

그림 10. 커널 사이즈 2×2의 정확도
Fig. 10. Accuracy of kernel size 2x2

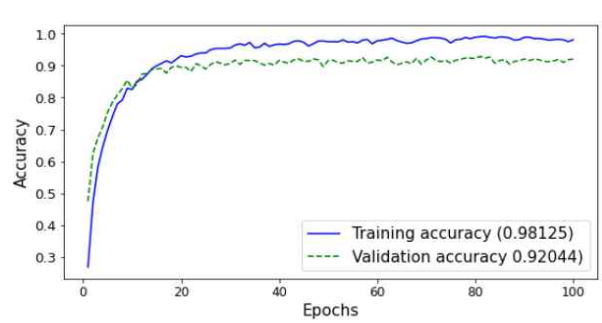

그림 11. 커널 사이즈 3×3의 정확도
Fig. 11. Accuracy of kernel size 3x3

Ⅵ. 결 론

본 논문의 실험에서는 오디오 분류 연구에서 다수 사용되는 UrbanSound8K 데이터 셋의 환경음 10 종류에서 MFCC를 추출하여 CNN에 학습할 때, 커널 사이즈와 풀링 레이어의 개수를 변경하며 분류 정확도를 관찰하였다. 이미지 분류와 다르게 오디오 분류 CNN에서는 waveform에서 추출한 MFCC 혹은 Mel spectrogram을 학습한다. 따라서 본 논문에서는 CNN을 통하여 오디오 분류를 수행할 때, 커널 사이즈를 크게 설정할수록 성능에 긍정적인 영향을 미치는 결과가 나타났다.

세 가지 시나리오에서 커널 사이즈와 풀링 레이어의 개수를 증가시킬수록 정확도가 증가하고 파라미터의 개수는 감소하는 결과가 도출되었다. 위 결과는 실험에서 사용된 Urbansound8K 데이터 셋과 유사한 환경음에서 MFCC를 추출하여 CNN으로 오디오 분류를 수행할 경우, 커널 사이즈 및 풀링 레이어의 개수의 결정에 활용할 수 있을 것으로 기대된다. 향후에 커널 사이즈와 풀링 레이어의 개수를 증가시킬수록 성능이 개선되는 이론적인 원인 분석을 진행할 예정이다.

참고문헌

[1] S. E. Mun, S. B. Jang, J. H. Lee, and J. S. Lee, “Trends of Machine Learning and Deep Learning Technology,” *Journal of The Korean Institute of Communication Sciences*, Vol. 33, No. 10, pp. 49-56, October 2016.

[2] M. O. Lee, U. N. Yoon, S. H. Ko, and G. S. Jo, “Efficient CNNs with Channel Attention and Group Convolution for Facial Expression Recognition,” *Journal of KIISE*, Vol. 46, No. 12, pp. 1241-1248, December 2019. https://doi.org/10.5626/JOK.2019.46.12.1241

[3] J. H. Kim and K. T. Choi, “Microcontroller-based Gesture Recognition using 1D CNN,” *Proceedings of the Korean Society of Computer Information Conference*, Vol. 29, No. 1, pp. 219-220, January 2021.

[4] M. S. Lee and H. C. Ahn, “A Time Series Graph Based Convolutional Neural Network Model for Effective Input Variable Pattern Learning : Application to the Prediction of Stock Market,” *Journal of Intelligence and Information System,* Vol. 24, No. 1, pp. 167–181, March 2018. https://doi.org/10.13088/jiis.2018.24.1.167

[5] W. H. Jo, Y. H. Lim, and K. H. Park, “Deep learning based Land Cover Classification Using Convolutional Neural Network - a case study of Korea,” *Journal of the Korean Geographical Society*, Vol. 54, No. 1, pp. 1-16, February 2019.

[6] J. B. Kong and M. S. Jang, “Association Analysis of Convolution Layer, Kernel and Accuracy in CNN,” *The Journal of the Korea institute of electronic communication sciences*, Vol. 14, No. 6, pp. 1153–1160, December 2019. https://doi.org/10.13067/JKIECS.2019.14.6.1153.

[7] W. G. Oh, “Comparison of environmental sound classification performance of convolutional neural networks according to audio preprocessing methods,” *The Journal of the Acoustical Society of Korea,* Vol. 39, No. 3, pp. 143–149, May 2020. https://doi.org/10.7776/ASK.2020.39.3.143

[8] M. K. Lim, D. H. Lee, H. S. Park, and J. H. Kim, “Audio Event Detection Using Deep Neural Networks,” *Journal of Digital Contents Society*, Vol. 18, No. 1, pp.183-190, February 2017. http://dx.doi.org/10.9728/dcs.2017.18.1.183

[9] Y. J. Park and H. S. Cho, “An Experiment of Sound Recognition using Machine Learning," *2020 IEEE International Conference on Consumer Electronics – Asia*, Seoul, pp. 1-3, December 2020. http://dx.doi.org/10.1109/ICCE-Asia49877.2020.9277368.

[10] M. K. Lim, D. H. Lee, K. H. Kim, and J. H. Kim, “Audio Event Classification Using Deep Neural Networks,” *Journal of the Korean Society of Speech Sciences*, Vol. 7, No. 4, pp. 27–33, December 2015. https://doi.org/10.13064/ksss.2015.7.4.027.

정윤아(Yun-A Jung)

2019년～현 재 : 강원대학교 IT대학 전기전자공학과 재학

※관심분야 : 인공지능(AI), 딥러닝, 심층 신경망, CNN 등

김동회(Dong-Hoi Kim)

2005년 : 고려대학교 전파공학과 (공학박사)

1989년 1월～1997년 1월: 삼성전자 전임연구원
2000년 8월～2005년 8월: 한국전자통신연구원 선임연구원
2006년 3월～현 재: 강원대학교 IT대학 전기전자공학과
2020년 6월～현 재: 강원대학교 정보화본부장 등
※관심분야 : 무선 네트워크 및 사물인터넷(IoT) 등