Journal of Knowledge Information Technology and Systems(JKITS), Vol. 15, No. 6, pp. 973-981, December 2020
https://doi.org/10.34163/jkits.2020.15.6.006

Journal of Knowledge Information Technology and Systems

ISSN 1975-7700 (Print), ISSN 2734-0570 (Online)

http://www.kkits.or.kr

Sound Classification Performance of Deep Neural Networks in the Presence of Disturbing Sounds

Wongeun Oh[*1], Dong-Kyun Lim[2]

[1]*Department of Multimedia Engineering, Sunchon National University*
[2]*Department of Computer Science Engineering, HanYang Cyber University*

ABSTRACT

Environmental sound classification is an area that automatically classifies sounds in our surroundings. It can be applied to home automation, security, and surveillance. Recently, the deep learning approaches have been adopted as a classifier for increasing performance. In this method, a deep neural network is trained using many sound data, and after the learning is completed, the microphone pickup sound is applied for classification. However, during this stage, the ambient noise can be put into the microphone along with the sound to be identified. And the sound cannot be properly classified due to this disturbing noise. The recognition rate of the deep neural network decreases as the loudness of the disturbing sound increases, but the analysis about the noise effect on the classification have been limited. In this paper, we present the effect of the disturbing noise on the classification rate. For this purpose, UrbanSound8K, which is composed of 10 types of urban environmental sounds, is used for training and test data. And the VGG16-based CNN which shows good performance in image classification was adopted as a baseline model. For the disturbing noise, we use three types of sounds that are consist of daily noises (hairdryer, vacuum cleaner, faucet water, and hammer), voices(male, female, and synthesized sound), and music(cello, piano, and trumpet). In the experiment, these disturbing noises are mixed with the clean sounds so that the signal-to-noise ratio is in the range of -50dB to 50dB. Then the mixed sound was applied to a deep neural network to obtain a relative recognition rate when compared to the clean cases. The results show that the recognition rate is more than 90% compared to the clean sound cases when the SNR is between 10 and 15 dB, and 95% or more when the SNR is greater than 20 dB, regardless of the type of the disturbing sound.

KEYWORDS: Environmental sound classification, Noise, CNN, VGG16, UrbanSound8K, Signal to noise ratio

ARTICLE INFO: Received 27 November 2020, Revised 9 December 2020, Accepted 11 December 2020.

[*]Corresponding author is with the Department of Multimedia Engineering, Sunchon National University, 255 Jungang-ro, Suncheon, Jeollanam-do 57922, KOREA.
E-mail address: owg@scnu.ac.kr

1. 서 론

환경음 분류(environmental sound classification)는 일상의 주변 환경에서 흔히 들을 수 있는 소리를 자동으로 인식하고 분류하는 분야로서, 사물 인터넷, 원격 감시, 홈오토메이션 등에 응용 시 시각이 미치지 못하는 영역의 데이터 처리에 효과적으로 응용될 수 있다. 또한 점차 늘어나는 노년층을 위한 청각 보조 기구 또는 청각 장애인을 위한 주변 상황 알림 등에도 유용하게 사용될 수 있다.

최근에는 이러한 환경음 분류 시 머신 러닝 기반의 심층 신경망을 이용하여 인간의 청각과 유사한 수준으로 인식률을 높이려는 연구가 활발하게 진행되고 있다[1-9]. 이 방법은 먼저 구분하려는 소리 데이터를 이용해서 심층 신경망을 학습한 다음, 학습이 완료된 신경망에 마이크 등으로 실제 소리를 입력하여 어떤 소리인지 구분하는 방식으로 동작한다. 그런데 이를 실제로 응용할 때 인식하려는 소리와 주변의 소리도 함께 입력된다. 예를 들어 가정에서는 청소기나 헤어드라이어와 같은 가전제품 작동음, 사람 목소리, 그리고 음악 소리 등과 같은 주변 소음이 판별하려는 소리와 함께 마이크로 입력되어 인식률에 영향을 준다. 이때 방해 소음의 크기가 커지면 심층 신경망의 인식률이 저하될 것으로 예상할 수 있으나, 소음의 종류와 레벨에 따라 구체적으로 어느 정도 영향을 주는가에 대한 자료는 찾아보기 힘들다.

본 논문에서는 머신 러닝을 이용하여 환경음 분류 문제를 다룰 때 방해음으로 인한 인식률 영향을 실험적으로 분석하였다. 이를 위해 먼저 UrbanSound8K 데이터셋[10]을 사용하여 VGG16[11]을 기반으로 한 심층 신경망을 학습시켰다. UrbanSound8K 데이터셋은 10종의 도시 환경음으로 구성되어 있으며, 다수의 연구에서 데이터셋으로 사용되었다[6-9]. VGG16은 이미지 분류에서 좋은 성능을 보이는 16층으로 구성된 CNN으로서, 실험에서는 가중치(weight)를 UrbanSound8K 데이터셋으로 다시 학습시켜 사용하였다.

방해음은 주거 공간의 일상 생활에서 흔히 들을 수 있는 생활 소음으로 선정하였으며, 가전제품 작동음(헤어드라이어, 청소기, 수돗물 소리), 남/여 목소리, 망치 소리, 음악(첼로, 피아노, 트럼펫)등으로 구성되어 있다. 실험은 먼저 방해음을 판별하려는 소리와 신호대 잡음비가 −50dB ~ 50dB가 되도록 합성한 다음, 이를 학습이 완료된 심층 신경망에 입력하여 방해음이 없을 경우 대비 인식률이 어느 정도 저하되는지 실험하였다. 실험 결과 SNR이 10dB~15dB인 경우 방해음이 없을 경우 대비 90% 이상의 인식률을 보였고, SNR이 20dB 이상일 때는 방해음의 종류와 무관하게 95% 이상의 상대 인식률을 나타내었다.

본 논문의 구성은 2장은 데이터셋과 신경망 구조를 다루고, 3장은 실험 결과 및 분석 그리고 4장은 결론으로 이루어져 있다.

2. 데이터셋과 신경망 구조

2.1 도시 환경음 데이터셋

본 논문에서 사용한 UrbanSound8K 데이터셋은 도시 일상에서 흔히 들을 수 있는 에어컨 소리, 개 짖는 소리, 자동차 경적 등과 같은 10종류의 소리로 구성되어 있다. 데이터셋의 전체 음원의 수는 총 8732개이며, 각 음원은 4초 이하 길이의 wav형식으로 저장되어 있다. 소리의 종류와 데이터의 수는 <표 1>과 같다.

2.2 오디오 데이터 특징 추출

CNN을 이용해서 오디오를 분류하기 위해서는

먼저 각 음원에서 특징을 추출해야 한다. 오디오의 특징 추출 방법은 여러 가지가 있으나 본 논문에서는 심층 신경망을 이용한 음향 분류 문제에서 타 특징에 비해 높은 인식률을 나타내는[12, 13] 로그 멜 스펙트로그램(log mel spectrogram)을 사용하였다. 추출 과정은 먼저 각 음원을 46.4ms 단위로 프레임을 구성하고 이를 50%씩 중첩하며 총 174개의 프레임을 생성한 다음, 각 프레임당 128개의 멜 밴드 에너지를 계산하여 128×174 크기의 로그 멜 스펙트로그램 데이터를 추출하였다.

표 1. UrbanSound8K 데이터셋의 소리 종류와 음원의 수
Table 1. The sound classes and number of audio clips in the UrbanSound8K dataset.

classID	Sound class (Abbreviation)	Number of clips
0	air_conditioner (ac)	1,000
1	car_horn (ch)	429
2	children_playing (cp)	1,000
3	dog_bark (db)	1,000
4	drilling (dr)	1,000
5	engine_idling (en)	1,000
6	gun_shot (gs)	374
7	jackhammer (jh)	1,000
8	siren (si)	929
9	street_music (st)	1,000

2.3 CNN 구조와 학습 방법

VGG16은 ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[14]에서 우수한 성능을 나타낸 16층으로 구성된 CNN구조의 심층 신경망으로서 이미지 분류 문제에서 자주 사용되는 모델이다. 실험에서는 VGG16의 출력단에 3개층의 relu 활성 함수와 1개층의 softmax 층을 추가하여 UrbanSound8K 데이터셋의 10개 소리를 구분하도록 하였다. 그리고 Keras에서 제공되는 VGG16 모델은 이미지넷의 데이터로 학습한 것이기 때문에 오디오 데이터에 그대로 사용할 수 없으므로 전체 네트워크의 가중치를 UrbanSound8K 데이터로 다시 학습시켰다. 학습 시 오디오의 로그 멜 스펙트럼은 VGG16에 적합한 224 x 224 x 3 크기의 jpg형식으로 변환하여 인가하였고, 학습 파라미터로 미니 배치 크기는 32, 학습률은 0.0001, 최적화 함수는 Adam 알고리즘을 사용하였다. 이때 전체 데이터의 10%는 검증용으로 사용하여 최대 100 에포크 동안 훈련하였으며, 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. 인식률은 UrbanSound8K데이터에서 권장하는 10-겹(10-fold) 교차 검증법을 사용하여 10회 평균 인식률을 구하였다.

3. 실험 절차

3.1 방해음 데이터 생성

방해음으로 사용할 소리는 일상에서 흔히 들을 수 있는 것으로 선택하였으며, 음원의 특성과 종류를 고려하여 다음과 같은 소리를 사용하였다.

- 생활 소음 : 헤어드라이어, 청소기, 싱크대 물 소리, 망치 소리
- 음성 : 남성, 여성, 합성음
- 음악 : 첼로, 피아노, 트럼펫

이들 음원에는 해당 소리 이외의 다른 소리는 포함되지 않았으며, 공개 음원 사이트인 freesound.org[15]에서 얻은 압축되지 않은 wave 포맷의 녹음 데이터(샘플링율 44.1kHz, 양자화 16비트 이상) 또는 고성능 마이크(Rode NTG4+)를 사용하여 직접 녹음한 소리를 사용하였다.

실험 데이터는 이들 방해음과 UrbanSound8K 음

원 각각을 다음 식 1과 같이 특정 SNR이 되도록 합성하여 생성하였다.

$$SNR = 10\log\frac{\sum_k s^2(k)}{\sum_k n^2(k)} \quad (1)$$

여기에서 $s(k)$는 방해음이 없는 UrbanSound8K의 소리이고, $n(k)$는 방해음이다. SNR은 -50dB ~ 50dB 범위에서 구했으며, 인식률이 급격하게 변하는 −20dB ~ 20dB 구간에서는 5dB간격으로 데이터를 생성하고 그외 구간에서는 10dB ~ 20dB간격으로 실험 음원을 생성하였다. 생성한 데이터는 방해음 15종, SNR 13개로서 총 1,702,740개로 구성된다.

3.2 기준 인식률

먼저 방해음이 없는 UrbanSound8K 데이터를 이용하여 소리를 분류하도록 심층 신경망을 학습시킨 다음 이때의 인식률을 기준 인식률로 사용한다.

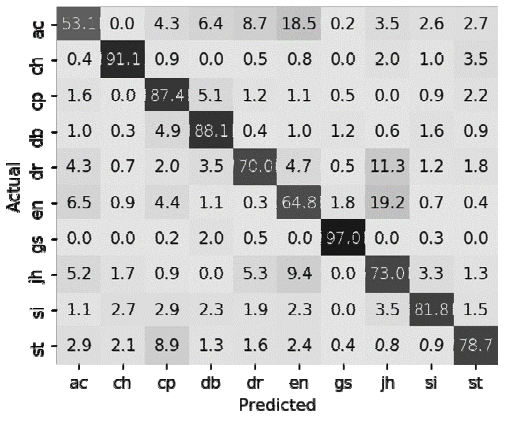

그림 1. 방해음이 없을 때 인식률의 혼동 행렬
Figure 1. The confusion matrix of deep neural network outputs without disturbance noise

기준 인식률 학습 시 10-fold 평균 인식률은 0.764(76.4%)이고, 각 음원별로는 최대 97% (gun_shot), 최저 53.1% (aircon)의 인식률을 나타내었다. <그림 1>은 방해음이 없을 때의 혼동 행렬(confusion matrix)을 나타낸 것이다.

4. 실험 결과

학습이 완료된 심층 신경망에 방해음이 섞인 소리를 입력하여 SNR별 인식률을 방해음이 없을 때와 비교하였다. 전체 결과를 <표 2>에 보였다. 표의 값은 아래 <식 2>와 같이 방해음이 있을 때의 인식률을 방해음이 없을 때의 학습률인 0.767로 정규화하여 상대적인 인식률을 %로 나타낸 것이다.

$$\text{Relative Accuracy}(\%) = \frac{\text{Accuracy}}{0.764} \times 100 \quad (2)$$

4.1 생활 소음의 영향

가정에서 흔히 듣는 소음으로 헤어 드라이어, 진공 청소기, 망치질 소리(0.8초에 1회), 그리고 수돗물 소리를 사용하였다. SNR에 따른 상대 인식률을 <그림 2>에 나타내었다.

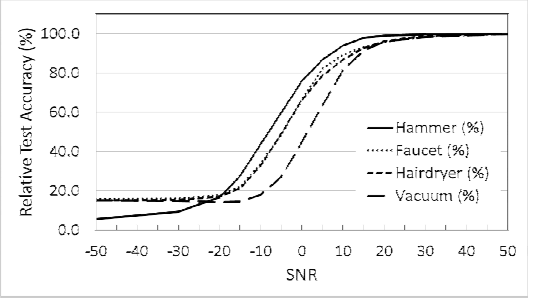

그림 2. 방해음이 생활 소음일 때 상대 인식률
Figure 2. The relative accuracy of deep neural network in the presence of daily noises

표 2. 방해음 SNR에 따른 상대 인식률(%)

Table 2. Relative test accuracy(%) in the presence of the disturbing noises

SNR(dB) / Noise	−50	−30	−20	−15	−10	−5	0	5	10	15	20	30	50
Daily noise													
Hairdryer	15.1	15.3	17.1	21.2	32.7	49.2	66.1	78.8	86.9	92.6	96.1	98.9	100.0
Vacuum	15.2	14.9	14.3	14.4	17.9	27.1	44.7	63.7	81.6	91.2	96.1	98.4	100.0
Hammer	5.7	9.3	16.6	27.7	43.9	59.8	76.0	86.8	93.9	97.9	98.9	99.7	100.0
Faucet	15.9	15.9	17.9	22.2	34.0	48.3	66.3	82.4	89.2	93.1	95.7	98.7	99.9
Daily noise (Avg.)	**13.0**	**13.9**	**16.5**	**21.4**	**32.1**	**46.1**	**63.3**	**77.9**	**87.9**	**93.7**	**96.7**	**98.9**	**100.0**
Voice													
Voice_male	15.8	16.6	20.3	25.1	35.2	53.6	71.7	85.5	93.7	97.4	99.5	100.2	100.0
Voice_male(syn)	15.4	16.1	19.3	23.2	32.8	49.0	67.1	81.1	91.5	96.4	98.7	99.9	99.9
Voice_female	15.4	15.7	19.9	25.7	32.4	42.6	57.0	72.3	84.5	91.9	96.3	99.8	100.1
Voice_female(syn)	15.9	16.6	21.0	26.1	34.1	46.8	62.1	77.5	88.5	95.3	98.7	100.0	99.9
Voice (Avg.)	**15.6**	**16.2**	**20.1**	**25.0**	**33.6**	**48.0**	**64.5**	**79.1**	**89.5**	**95.2**	**98.3**	**100.0**	**100.0**
Instruments													
Cello	15.4	16.7	19.5	23.7	31.8	46.2	62.8	79.0	91.1	97.0	98.9	99.7	99.9
Piano	13.8	14.8	16.4	19.1	22.9	30.1	42.7	60.3	78.7	90.1	95.7	99.5	100.0
Piano_jazz	15.6	15.6	15.7	16.1	18.0	21.4	30.4	51.1	76.4	90.5	95.9	99.3	99.9
Trumpet	9.7	12.2	14.5	16.6	20.4	26.9	36.3	49.1	62.9	76.5	86.8	97.5	100.0
Instruments (Avg.)	**13.6**	**14.8**	**16.5**	**18.9**	**23.3**	**31.2**	**43.0**	**59.9**	**77.3**	**88.5**	**94.3**	**99.0**	**99.9**
Total Avg.	**14.1**	**15.0**	**17.7**	**21.8**	**29.7**	**41.7**	**56.9**	**72.3**	**84.9**	**92.5**	**96.4**	**99.3**	**100.0**

인식률에 가장 영향이 적은 것은 망치 소리였으며, 물소리와 헤어드라이어는 비슷하며 진공 청소기 소리는 상대적으로 가장 높은 방해 요인으로 나타났다.

인간이 청감적으로 2배 크기로 인식되는 SNR 10dB에서 생활 소음의 상대 인식률 평균은 87.9%였으며, 가장 높은 인식률을 보인 망치 소리와 가장 낮은 청소기 소리의 차이는 12.3%로 방해음의 종류에 따라 인식률 차이가 크다는 것을 알 수 있다. 방해음이 없는 경우와 거의 동등한 성능을 내기 위해서는 4종의 방해음 모두 SNR이 30dB 이상이 되어야 하며, 이때 기준 대비 98% 이상의 인식률을 나타내었다.

4.2 음성의 영향

방해음이 음성일 때의 인식률 영향을 실험하기 위해 남성, 여성, 남성 기계 합성음, 여성 기계 합성음으로 동일한 문장을 녹음하여 실험하였다. <그림 3>은 음성 방해음의 상대 인식률 결과를 나타낸 것이다. 가장 영향이 적은 것은 남성 음성이었고 다음으로 남성 합성음, 여성 음성, 그리고 여성 합성음 순으로 나타났다.

SNR=10dB 일 때 상대 인식률은 평균 89.5%로 생활 소음에 비해 방해 정도가 덜한 것으로 나타났으며, 최대 상대 인식률과 최소 상대 인식률 간의 편차가 9.2%로 생활 소음보다 낮게 나타났다. 또한 남성 음성이 여성 음성보다 방해 정도가 낮게 나타났다. 예를 들어 SNR=10dB에서 남성 음성과 여성 음성의 상대 인식률 평균은 각각 92.6%와 86.5%로 남성 음성이 약 6.1% 더 높았다.

4.3 악기음에 의한 영향

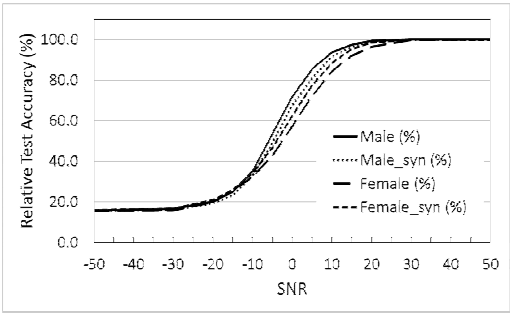

그림 3. 방해음이 음성일 때 상대 인식률
Figure 3. The relative accuracy of deep neural network in the presence of voices

악기음을 평가하기 위해서 현악기(첼로), 관악기(트럼펫), 그리고 건반악기(피아노) 3종의 악기에 대해 각각 멜로디 연주와 음계 연주일 때에 대해 실험하였다. <그림 4>는 그 결과를 나타낸 것이다.

악기음 중 가장 영향이 가장 낮은 것은 첼로 였고 다음으로 피아노, 트럼펫 순으로 나타났다.

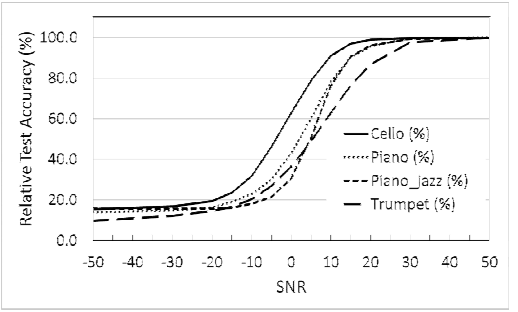

그림 4. 방해음이 악기음일 때 상대 인식률
Figure 4. The relative accuracy of deep neural network in the presence of musical instrument sound

피아노는 음악의 종류에 따라 SNR 5dB 이하에서 차이를 보였지만 SNR이 10dB 이상인 경우에는 거의 동일한 인식률을 보였다. 10dB에서 평균 상대 인식률은 77.3% 였으며, 최대 인식률과 최저 인식률의 차이는 29%(첼로 91.9%, 트럼펫의 62.9%)의 차이를 보여 세 가지 방해음 유형 중에서 가장 큰 편차를 나타냈다. 이는 어느 정도 스펙트럼이 제한된 생활 소음이나 음성과는 달리 악기 종류에 따라 스펙트럼 대역이 차이가 나기 때문이며, 고음을 내는 악기가 방해의 정도가 큰 경향을 나타냈다.

4.4 방해음 유형별 영향

<그림 5>는 SNR에 따른 각 방해음 유형별 상대 인식률 평균을 나타낸 것이다. 소음 유형(일상 소음, 음성, 악기음)별로 방해 정도는 음성이 평균적으로 가장 낮았고, 다음으로 생활 소음, 악기의 순서였다. 음성과 생활 소음의 차이는 대략 1~2% 정도로 그렇게 크지 않았으나, 악기음은 악기 종류에 따라 일상 소음 대비 최대 20%까지도 차이가 나는 경우가 있어서 악기 종류에 따라 방해의 정도가 크게 달라지는 것을 알 수 있다.

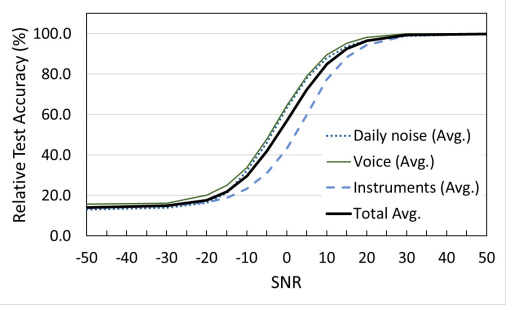

그림 5. 방해음 유형별 상대 인식률
Figure 5. The relative accuracy of deep neural network according to the disturbance type

<표 3>은 방해음 없을 때 대비 90% 인식률과 95% 인식률을 얻을 수 있는 SNR값을 유형별로 나타낸 것이다. 일상 소음이나 음성만 존재하는 경우에는 SNR이 11.8dB 이상이면 무소음 대비 90%의 상대 인식률을 얻을 수 있으며, 17.2dB 이상이면 95%의 상대 인식률이 얻어진다. 그러나 악기음이 존재하는 경우는 90%의 상대 인식률을 얻기 위해

서는 16.3dB, 95%의 상대 인식률을 얻기 위해서는 21.5dB의 SNR이 필요함을 알 수 있다.

표 3. 90%와 95% 상대 인식률에 필요한 SNR
Table 2. The SNR for 90% and 95% relative accuracy

Noise \ Rel. Accu.	90%	95%
Daily noise	11.8 dB	17.2 dB
Voice	10.5 dB	14.8 dB
Instrument	16.3 dB	21.5 dB
Total	13.3 dB	18.2 dB

5. 논의 및 결론

본 논문에서는 도시 환경음을 인식하도록 학습된 심층 신경망을 실제로 응용하는 경우 주변 소음으로 인해 인식률이 저하되는 정도를 여러 유형의 방해음에 대해 분석하였다. 실험 결과 동일한 SNR인 경우에 음성이 가장 낮은 방해 요인이었으며, 다음으로 생활 소음, 악기음 순이었다. 음성과 생활 소음은 종류에 따라 큰 차이가 없었다. 그러나 악기음은 악기의 종류에 따라 인식률에 많은 차이가 발생하였으며, 첼로와 같은 저음 악기보다 트럼펫과 같은 고음 악기가 인식률에 더 많은 영향을 주었다. 음성의 경우 남성 음성보다 여성 음성으로 인한 방해 정도가 약간 더 높게 나타났다. 대체적으로 저음역대 보다는 중고 음역대의 방해음이 인식률을 낮추는데 더 많은 영향을 주는 경향이 있었다.

이상의 연구 결과로 어떤 소리가 환경음 인식 시 더 큰 방해 요인으로 작용하는지에 대해 알 수 있으며, 이를 방해음 크기와 인식률에 대한 가이드라인으로 사용할 수 있을 것이다. 그러나 본 연구는 특정 구조의 심층 신경망을 이용해서 제한된 종류의 방해음을 이용하여 인식률을 평가한 것으로 일반화에는 한계가 있으므로, 보다 일반적인 결과를 도출하기 위해 순환 신경망과 같은 또 다른 구조의 심층 신경망과 다른 종류의 데이터셋, 그리고 보다 다양한 방해음을 이용한 실험을 수행할 예정이다.

References

[1] S. Abdoli, P. Cardinal, and A. L. Koerich, *End-to-end environmental sound classification using a 1D convolutional neural network*, Expert Systems with Applications, Vol. 136, pp. 252-263, 2019.

[2] B. Zhu, C. Wang, F. Liu, J. Lei, Z. Huang, Y. Peng, and F. Li, *Learning environmental sounds with multi-scale convolutional neural network*, Proceedings of the International Joint Conference on Neural Networks, pp. 1-8, 2018.

[3] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, *Classifying environmental sounds using image recognition networks, Procedia Computer Science*, Vol. 112, pp. 2048-2056, 2017.

[4] Y. Tokozume and T. Harada, *Learning environmental sounds with end-to-end convolutional neural network*, 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2721-2725, 2017.

[5] D. Chong, Y. Zou, and W. Wang, *Multi-channel convolutional neural networks with multi-level feature fusion for environmental sound classification*, Lecture Notes in Computer Science, pp. 157-168, 2019.

[6] K. J. Piczak, *Environmental sound classification with convolutional neural*

networks, Proceeding of IEEE 25th International Workshop on Machine Learning for Signal Processing, pp. 1-6, 2015.

[7] Y. Su, K. Zhang, J. Wang, and K. Madani, *Environment sound classification using a two-stream CNN based on decision-level fusion*, Sensors, Vol. 19, pp. 1733-1747, 2019.

[8] S-M. Suh, *Effective implementation for fast deep learning algorithm*, Journal of Knowledge Information Technology and Systems, Vol. 14, No. 5, pp. 553-561, 2019

[9] M-J. Kang, Y-S. Kim, H-Y. Shin, J. Park, *Development of the deep learning system for bird classification using birdsong*, Journal of Knowledge Information Technology and Systems, Vol. 15, No. 2, pp. 195-203, 2020

[10] 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, 2014.

[11] K. Simonyan and A. Zisseman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, 2015.

[12] W. Oh, *Audio classification performance of CNN according to audio feature extraction methods*, Proceedings of Acoustical Society of Korea, p. 64, 2019.

[13] W. 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, 2020.

[14] 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*, International Journal of Computer Vision, Vol. 115, pp. 211-252, 2015.

[15] Freesound, https://freesound.org/, Oct. 2020.

딥러닝을 이용한 소리 분류 시 방해음의 영향 분석

오원근[1], 임동균[2]

[1]*순천대학교 멀티미디어공학전공 교수*
[2]*한양사이버대학교 컴퓨터공학과 교수*

요 약

환경음 분류는 일상 환경에서 흔히 들을 수 있는 소리를 자동으로 인식해서 분류하는 분야이다. 이는 가정, 보안, 감시 등 분야에서 시각이 미치지 못하는 영역의 데이터 처리에 유용하게 사용될 수 있어 점차 관심이 커지고 있으며, 최근에는 CNN과 같은 심층 신경망을 이용하여 인식률을 높이려는 연구가 활발하게 진행되고 있다. 이 방법에서는 먼저 대량의 소리 데이터를 이용하여 심층 신경망을 학습한 후, 학습이 완료된 신경망에 마이크 등으로 소리를 입력하는 방식으로 동작한다. 그런데 응용하는 단계에서 인식하려는 소리와 함께 주변 소음이 판별하려는 소리와 함께 마이크로 입력되며, 이러한 방해음으로 인해 원래 학습했던 소리의 분류 성능이 떨어지게 된다. 이때 방해음의 크기가 커질 수록 심층 신경망의 인식률이 저하될 것으로 예상할 수 있으나, 구체적인 소음의 종류와 크기가 인식률에 미치는 영향에 대한 분석은 찾아보기 힘들다. 본 논문에서는 CNN을 이용하여 환경음 분류 시 방해음으로 인한 인식률 영향을 실험적으로 분석하였다. 사용한 데이터셋은 10종의 도시 환경음으로 구성된 UrbanSound8K이며, 심층 신경망으로는 이미지 분류에서 좋은 성능을 보이고 있는 VGG16 기반의 CNN을 이용하였다. 방해음은 주거 공간의 일상에서 흔히 들을 수 있는 3가지 유형의 소리를 사용하였는데, 생활 소음(헤어드라이어, 청소기, 수돗물, 망치),

음성(남자, 여자, 합성음), 음악(첼로, 피아노, 트럼펫) 등으로 구성되어 있다. 실험은 판별하려는 소리와 방해음을 신호대 잡음비가 −50dB ~ 50dB 범위가 되도록 합성한 다음, 이를 학습이 완료된 심층 신경망에 입력하여 방해음이 없을 경우와 비교한 상대적 인식률을 구하였다. 실험 결과 SNR이 10~15dB 구간에서는 방해음이 없을 경우와 비교 시 90% 이상의 상대 인식률을 보였고, SNR이 20dB 이상인 경우는 방해음의 종류와 무관하게 95% 이상의 상대 인식률을 나타내었다.

감사의 글

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

Wongeun Oh received the B.S., M.S., and Ph.D. degrees in Electronic Communication Engineering from Hanyang University in 1989, 1992, and 1997, respectively. In 1997, he joined the faculty member of Sunchon National University, where he is currently a professor in Department of Multimedia Engineering. His current research interests include sound systems, machine learning, and audio signal processing.

E-mail address: owg@scnu.ac.kr

Dong Kyun Lim received the bachelor's degree in the Department of Electronic Communication Engineering from Hanyang University in 1985. He received the M.S. degree and the Ph.D. degree in the Department of Electronic Communication Engineering from Hanyang University in 1987 and 2001, respectively. He was a professor in the Department of Computer Science Engineering at ChungCheng University from 1990 to 2003. He has been a professor in the Department of Computer Science Engineering at HanYang Cyber University since 2003. His current research interests include artificial intelligence, cyber education, microprocessor. He is a life member of the KKITS.

E-mail address: eiger07@hycu.ac.kr