|  |
| --- |
| **CNN기반의 주변환경 음향 구분** |
|  |
| Environment Sound Classification Based On CNN |
|  |
| **요 약**  본 논문은 주변 환경에서 발생되는 다양한 음향의 분류를 통해 장소를 특정하는 방법에 관한 논문이다. 본 논문에서 사용된 분류 데이터는 장소를 특정하여 주변의 음향 데이터를 수집하였고, Convolution Neural Network(CNN) 모델을 기반으로 음향 환경 분류를 진행하였다. 실험은 크게 5가지 Category를 나누었고, 각 Category당 3개의 Sub-class를 나누어 총 15개의 Sub-class를 분류하여 정확도를 비교, 분석하였다. | | |

1. **서 론**

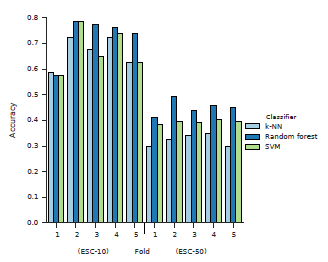
본 연구에서는 컴퓨터 공학 분야에서 화두가 되고 있는 인공지능(AI)을 음향 분야에 이용한 주제를 다루고자 한다. 인공지능의 급격한 발전은 컴퓨터 공학의 전통 분야에서 AI기법을 적용할 수 있는지를 논의함으로써 이루어졌다. 음성, 이미지, 자연어, 게임 등 우리가 접할 수 있는 데이터 전반에 걸쳐 최근까지도 적용 가능한 분야는 계속 증가하고 있는 추세이다.

인공지능 여러 분야 중 본 연구에서는 CNN(Convolution Neural Network) 모델을 기반으로 Environment Sound Classification(환경음향 분류)에 대한 연구를 진행한다. 먼저, MFCC(Mel-Frequency Cepstral Coefficient)로 Feature를 추출하고 CNN 기반의 모델을 적용하여 Spectrogram 기반의 환경 음향을 분류하게 되면 각각의 환경에서 음향의 주파수 에너지 변조 패턴을 파악할 수 있고, small receptive field를 가지는 convolution filter를 사용하여 네트워크를 훈련시킬 수 있기 때문에 각각의 환경을 특정할 수 있는 주파수 특징을 학습할 수 있어 CNN을 기반으로 한 모델은 음향 환경 구분의 인지에 있어 적합한 네트워크이다.[1]

1. **기존 연구**
   1. **ESC: Dataset for Environment Sound Classification**

이 논문은, 환경음향 데이터를 여러 Category를 분류하고 학습을 하여 그에 대한 결과와 사용한 CNN 모델을 제시한다.

ESC-50 데이터셋은 2000개의 labeled 환경음향으로 구성되어 있고 총 5개의 카테고리와 각 카테고리마다 10개의 클래스로 나누어져 있다.



**[그림 1 Comparison of classification accuracy between folds depending on the choice of classifier]**

[그림 1]에서 확인할 수 있듯이 카테고리가 많아 질수록 정확도가 낮아지는 것을 확인할 수 있다. 그리고 분류된 환경음향을 실제 4000명의 참여자들에게 들려주고 정확한 labeled 데이터셋을 고르는 실험을 했으며, 10개의 카테고리로 나눈 음향 데이터에서는 95.7%, 50개의 카테고리로 나눈 음향 데이터에서는 81.3%의 결과를 보였다. 특히 세탁기와 같이 모호한 특징을 가진 소리에 대해서는 34.1%, 뚜렷한 특징을 가진 아기 울음소리는 100% 분류했다.[2]

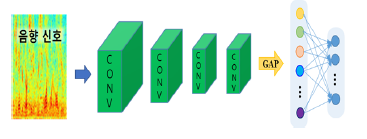
이 논문을 통해서 모호하고 다양한 노이즈가 섞여 있는 소리에서는 사람조차 분류하기 힘들다는 것을

밝히고 있다. 복합적인 노이즈가 섞여 있는 환경음향에 대해서 데이터를 어디서 수집하고 어떤 특징들을 추출할지 명확한 기준을 나누면 높은 정확도를 도출하는데 도움이 될 것이다.

* 1. **도시 환경에서의 환경음향 분류를 위한 Convolution Neural Network에 관한 연구 (A Study of the Convolutional Neural Network Structure for Urban Sound Classification)**

앞에서 제시한바와 같이 CNN은 환경음향을 인지하는 것에 있어 적합한 네트워크이다. 하지만 일반적인 convolution layer, pooling layer, fully-connected neural network로 구성된 CNN은 실제 주변 환경에서 발생되는 임의의 시간 길이를 가지는 환경음향 데이터를 적용하기 어렵다.[3]

이 논문은 그러한 한계를 보완하여 환경음향간의 시간 변동성을 고려하고 CNN의 localization능력을 살리기 위한 CNN구조를 제시한다.



**[그림 2] 제안된 환경 음향 분류 CNN구조**

[그림 2]는 앞에서 설명한 일반적인 CNN구조를 변경하여 뒷단의 fully-connected된 neural network를 fully convolution network, global pooling, softmax layer로 대체한 CNN구조를 이용하여 환경음향 분류 모델의 전체적인 네트워크 구조를 나타낸다.

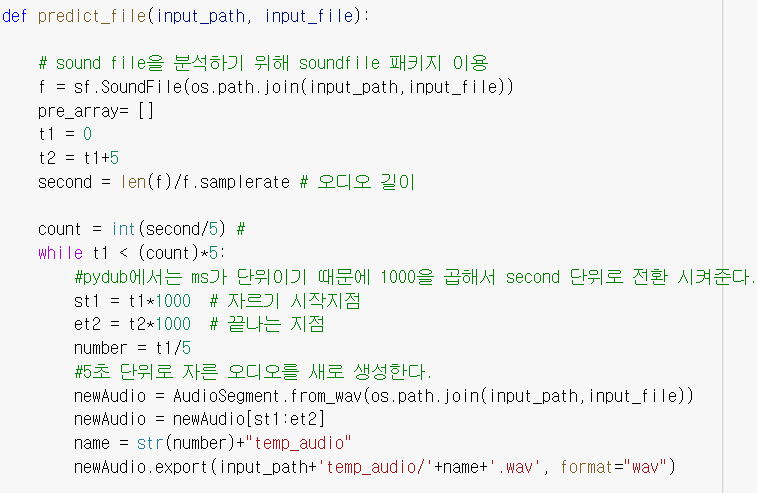
Fully convolution network는 convolution을 수행하는 network이며 기존의 fully-connected된 neural network의 역할을 대신한다. Global Average Pooling(GAP)은 마지막 convolution network의 feature map에 적용하여 정규화 기능을 수행하여 과적합을 줄여주는 역할을 수행한다.[4]

위 모델을 적용하면 각각의 filter당 하나의 특징을 얻을 수 있기 때문에 다양한 시간 길이를 가지는 환경음향들을 사용할 수 있다. 또한 성능 부분에서도 fully connected neural network를 가진 CNN보다 인식 성능이 2.8%정도 좋은 결과가 나왔다.

1. **시스템 모델**
   1. **기존 연구와 차이점 및 해결방안**

기존의 연구 및 논문에서는 환경음향 데이터로 특징이 뚜렷한 특징을 갖는 데이터를 사용하여 학습을 시켰다. 예를 들면 cat meowing, helicopter, brushing teeth와 같이 한가지 특징만을 갖고 있는 단일 음향으로 학습을 시키기 때문에 정확도가 보장이 된다. 본 연구에서는 길거리, 공항, 강의실, 카페 같이 여러가지 특징들이 섞여 있는 복합 음향을 학습시키고 새로운 장소에 대한 음향 데이터를 입력 데이터로 넣었을 때 장소를 구분하는 모델을 구현한다는 것이 기존 연구와의 차이점이다.

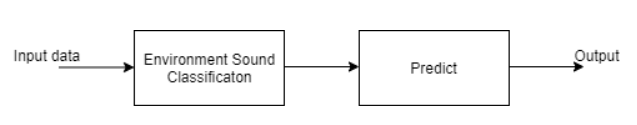
여러 복합 음향이 섞여 특정 장소는 객관적인 특징을 잡아내기가 어려울 것이라고 판단된다. 따라서 가능한 여러 장소의 음향을 수집하여 학습 데이터로 이용한다. 또한 환경음향간의 시간 변동성을 고려한 모델을 구축하여 임의의 시간 길이를 가지는 환경음향 데이터를 적용할 수 있도록 진행한다.



**[그림 3] Predict Model Code**

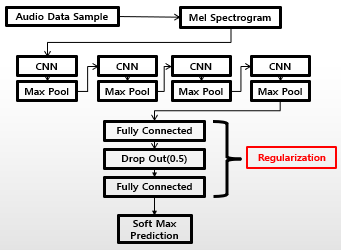
본 연구에서는 임의의 시간 길이를 가지는 환경음향 데이터를 넣었을 때 예측을 하는 모델을 구현했다. 데이터의 길이를 5초 단위로 나누어서 새로운 오디오 파일을 만들고 이를 librosa의 melspectrogram을 이용하여 모델의 입력에 맞는 tensor로 변환했다. 추출한 spectrogram을 model로 예측한 다음 flatten 합을 이용하여 일차원 배열 형태로 만들었다. 그 결과물로 나온 softmax중 가장 큰 값을 가지는 위치의 인덱스 값의 빈도수를 계산하여 가장 많은 빈도수가 나온 class를 찾아 입력 사운드에 대한 분류를 할 수 있게 했다.

**3.2 프로젝트 내용**



**[그림 4] 간략히 나타낸 전체적인 시스템 구조**

이 프로젝트는 여러 특징을 가진 복합 음향이 섞여 있는 입력 데이터를 학습된 모델을 통해서 장소를 구분하는 시스템이다. 모델의 구조는 아래와 같다.



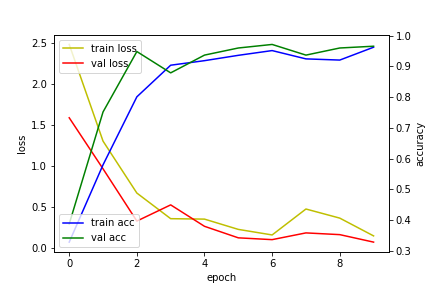
**[그림 5] Model Architecture**

오디오 데이터는 Sampling rate 44.1 kHz, Channel 1, Audio format은 WAV format으로 규격 한다. 규격화된 오디오 데이터 샘플을 Mel Spectrogram으로 변환하고 90%를 Train, 10%를 Validation, 학습에 적용되지 않은 10%분량의 데이터를 Test에 이용하며 Convolution layer, pooling layer, fully-connected neural network를 거쳐 Softmax activation function을 통해 특징을 학습시킨다.

학습된 모델의 Accuracy와 Loss를 확인하고 신뢰도가 낮다고 판단되면 데이터를 추가하거나 CNN구조를 변경을 하도록 한다.

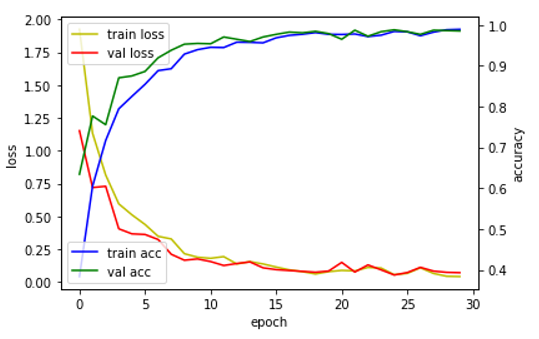
1. **결과**

다음은 실험을 통해 획득한 결과물과 그에 대한 분석이다.



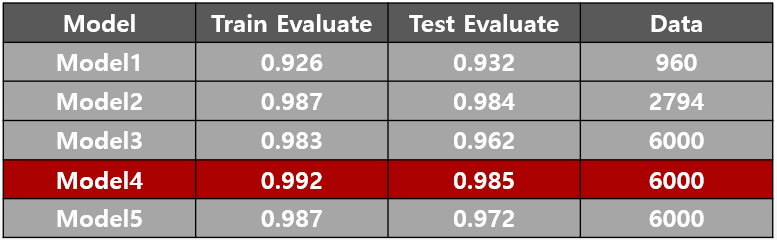
**[그림 6] Model1**

가장 처음에 했던 실험 결과로서, 3개의 카테고리와 9개의 서브 클래스로 분류하고 원본 자료 960개, 백색 소음을 추가로 생성하여 총 1920개의 데이터로 훈련 val비율 0.1로 하여 keras 모델로 학습을 진행했다. [그림 6]을 보면 2번의 epoch만에 accuracy가 크게 증가했고, 특히 epoch3에서 train acc가 val acc를 역전하는 현상이 발생해 과적합이 일어난 점을 확인할 수 있었다. 이 원인을 학습 데이터의 다양성 부족과 절대적인 값의 부족이라고 생각이 들어서 데이터를 추가적으로 생성해 훈련을 진행했다.



**[그림 7] Model4**

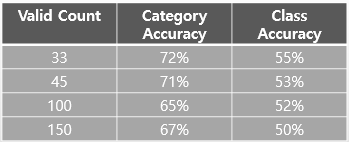
이후 동일한 CNN 모델에서 총 15개의 서브 클래스로 늘리고 총 입력 sample의 값을 2974(Train2)와 6000(Train3)개로 늘렸고 epoch를 30으로 늘리면서 학습 횟수를 증가시켰다. 그 결과 훈련 자체 내에서는 과적합이 일어나는 현상이 잘 보이지 않았지만 train data나 val data에 있는 값이 아닌 전혀 새로운 데이터를 넣었을 때 예측을 잘 못하는 경우가 발생하였다.



**[표 1]Model Evaluate**

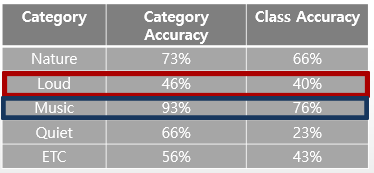
Model4 부터는 모델의 구조를 변경하였고 Fully connected의 크기를 (256, 22)에서 22를 제외하였고 Train5에서는 (256, 128)로 변경한 CNN을 이용했다. 그 결과 [그림 7]과 [표 1]에서 확인할 수 있듯이 Model4의 모델이 가장 loss 값이 일정하게 떨어지고 최종 accuracy가 가장 높게 나오는 모델이라는 것을 확인할 수 있다.

이 모델이 실제로 잘 작동하는지 테스트를 진행하였고, 기존에 모델 학습에 쓰이지 않은 전혀 새로운 data 150개를 30초로 변환하여 사용했다.



**[표 2]Model Valid Accuracy**

최종 accuracy는 Category 67%, Class 50%로 기대에 미치지 못한 결과를 보였고, 이에 대한 원인을 하나의 장소에 너무 많은 특징이 포함되어 큰 기준의 카테고리의 특징까지는 어느 정도 구분을 할 수 있지만 세부적인 영역에 있어서는 명확한 특징이 존재하지 않는 다면 구분을 하지 못한다는 결론을 내렸고 이를 증명하기 위해 분석을 하였다.



**[표 3]Category Accuracy**

[표 3]은 카테고리마다 정확도를 나타낸다. Music 카테고리에서 음악이라는 공통점을 가지고 있는 특징을

가지고 있는 원본 파일의 경우는 높은 확률로 music을 추측하게 된다. 하지만 음악 공연장에서 녹음한 소리이지만 음악적 특징이 적고 다른 특징이 많이 들어 있는 소리의 경우는 추측을 잘 하지 못했다. 때문에 큰 카테고리 영역에서는 구분을 어느 정도 하지만 세부적인 영역의 경우는 뚜렷한 특징을 가진 소리 구간이 아니면 구분을 제대로 못한다는 것이다.

실제로 사람들이 대화한 소리가 카페에서 대화하는 것인지 아니면 강의실에서 대화하는 것인지 생각보다 사람도 구별하는 것이 쉽지 않다. 하지만 그럼에도 이번 연구를 통해서 공통적인 커다란 카테고리를 영역을 구분하는 것은 실현이 가능하다는 점을 확인할 수 있었으며, 더 많은 특징을 가진 데이터를 확보하고 학습을 시킨다면 높은 정확도 얻을 수 있을 것이다.

1. **결론 및 향후 연구**

본 논문에서는 다양한 환경음향에서 음향에 대한 장소를 높은 정확도로 구분할 수 있는 모델을 구현하려고 한다. 일반화 성능을 위해서 더 많은 데이터를 수집하고 분류를 위한 명확한 Environment Sound Class를 정의하며 Mel Spectrogram 말고도 다른 음성 특징을 이용하여 연구하면 Valid Accuracy를 높일 수 있을 것이라고 기대된다.

AI를 이용한 음향 분야의 수요의 증가와 대중의 관심으로 AI 음성 인식의 시장 가치가 증가하고 있다. 따라서 AI 음향 분야는 여러가지 방향으로 계속 발전하고 소리 구분 및 인식 등의 기술도 계속 발전할 것으로 보인다. 소리 구분 및 인식의 활용 방안은 여러 범죄 상황이나 위험 감지 등 모니터링/감시, 청각장애인을 위한 sound visualization, 일상생활 모니터링과 같은 보조 기술 등을 예로 들 수 있다. 따라서 향후 연구를 통해 Environment Sound Classification을 Mobile과 IOT를 연계하여 확장된 Application을 실생활에 적용하는 방향으로 연구 및 개발을 진행한다.

* **후 주**

[1] 이윤진(2017), 「도시 환경에서의 환경음향 분류를 위한 Convolutional Neural Network에 관한 연구」, 『한국통신학회 학술대회논문집』, p.957.

[2] Karol J.Piczak, “ESC: Dataset for Environmental Sound Classification”, Institute of Electronic Systems Warsaw University of Technology, 2015, pp.2~4.

[3] Nac-Woo Kim(2019), 「Performance Comparison of Acoustic Features for Sound Classification on Ubiquitous Environment」, 『한국통신학회 학술대회논문집』, pp.1298~1299.

[4] 이윤진, Loc. cit..

* **참 고 문 헌**

김경수,「대화록 자동화를 위한 딥러닝 기반 한국어 화자분리방법」,건국대학교 ,건국대학교 석사학위, 2018.

오일석,『패턴인식』, 교보문고, 2018.

융합연구정책센터, 『융합연구리뷰』vol.3 no.6, 한국과학기술연구원 융합연구정책센터, 2017.

황일규, 「인공지능 기반의 울음소리를 이용한 영아 상태 인식」, 『한국정보기술학회논문지』, 한국정보기술학회, 2019.

이윤진,「도시 환경에서의 환경음향 분류를 위한 Convolutional Neural Network에 관한 연구」, 『한국통신학회 학술대회논문집』, 2017.

Karol J.Piczak, “ESC: Dataset for Environmental Sound Classification”, Institute of Electronic Systems Warsaw University of Technology, 2015.

Nac-Woo Kim, 「Performance Comparison of Acoustic Features for Sound Classification on Ubiquitous Environment」, 『한국통신학회 학술대회논문집』, 2019.

Tensorflow, <Speech Commands>,

<https://github.com/tensorflow/tensorflow/tree/master/tensorflow/examples/speech_commands> (2018.09. 29. 검색).