특집논문 (Special Paper)
방송공학회논문지 제25권 제6호, 2020년 11월 (JBE Vol. 25, No. 6, November 2020)
https://doi.org/10.5909/JBE.2020.25.6.854
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

심층 신경망을 통한 자연 소리 분류를 위한 최적의 데이터 증대 방법 탐색

박 진 배[a)], Teerath Kumar[a)], 배 성 호[a)‡]

Search for Optimal Data Augmentation Policy for Environmental Sound Classification with Deep Neural Networks

Jinbae Park[a)], Teerath Kumar[a)], and Sung-Ho Bae[a)‡]

요 약

심층 신경망은 영상 분류 그리고 음성 인식 등 다양한 분야에서 뛰어난 성능을 보여주었다. 그 중에서 데이터 증대를 통해 생성된 다양한 데이터는 신경망의 성능을 향상하게 시키는 데 중요한 역할을 했다. 일반적으로 데이터의 변형을 통한 증대는 신경망이 다채로운 예시를 접하고 더 일반적으로 학습되는 것을 가능하게 했다. 기존의 영상 분야에서는 신경망 성능 향상을 위해 새로운 증대 방법을 제시할 뿐만 아니라 데이터와 신경망의 구조에 따라 변화할 수 있는 최적의 데이터 증대 방법의 탐색 방법을 제안해왔다. 본 논문은 이에 영감을 받아 음향 분야에서 최적의 데이터 증대 방법을 탐색하는 것을 목표로 한다. 잡음 추가, 음의 높낮이 변경 혹은 재생 속도를 조절하는 등의 증대 방법들을 다양하게 조합하는 실험을 통해 경험적으로 어떤 증대 방법이 가장 효과적인지 탐색했다. 결과적으로 자연 음향 데이터 세트 (ESC-50)에 최적화된 데이터 증대 방법을 적용함으로써 분류 정확도를 향상하게 시킬 수 있었다.

Abstract

Deep neural networks have shown remarkable performance in various areas, including image classification and speech recognition. The variety of data generated by augmentation plays an important role in improving the performance of the neural network. The transformation of data in the augmentation process makes it possible for neural networks to be learned more generally through more diverse forms. In the traditional field of image process, not only new augmentation methods have been proposed for improving the performance, but also exploring methods for an optimal augmentation policy that can be changed according to the dataset and structure of networks. Inspired by the prior work, this paper aims to explore to search for an optimal augmentation policy in the field of sound data. We carried out many experiments randomly combining various augmentation methods such as adding noise, pitch shift, or time stretch to empirically search which combination is most effective. As a result, by applying the optimal data augmentation policy we achieve the improved classification accuracy on the environmental sound classification dataset (ESC-50).

Keyword: Convolutional neural networks, Environmental sound classification, Data augmentation, Searched augmentation

Ⅰ. 서 론

심층 신경망의 학습에서 데이터는 매우 중요한 역할을 한다. 훈련에 사용되는 데이터가 충분하지 않을 경우, 여기에 훈련된 신경망은 일반화되지 않은 편향된 예측을 할 확률이 높아지고 실생활에서 활용하기 어려워질 수 있다[1]. 데이터 개수의 부족을 보완하기 위해 데이터 증대 방법이 사용된다. 데이터 증대는 학습을 위해 수집한 데이터의 변형을 통해 더욱 다양한 형태로 만들어서 이를 통해 학습된 신경망이 더욱 일반화된 예측을 할 수 있도록 도와주는 기법이다[1,2,3].

최근 영상 처리 분야에서 객체 인식, 검출, 및 분할 등 다양한 작업에 사용되는 심층 신경망의 성능을 높이기 위해 데이터 증대에 관련된 연구도 활발히 진행되고 있다[2,3,4]. 영상의 좌우 반전, 회전, 혹은 명암 효과 등 다양한 변형을 시도할 뿐만 아니라 이들의 조합 개수 및 순서에 따른 영향도 연구되고 있다. 더 나아가서는 학습에 사용되는 데이터와 심층 신경망의 특성에 따라 최적의 데이터 증대 방법 및 조합이 다르다는 것을 실험을 통해서 증명하였으며, 각각의 상황에서 최적의 증대 방법을 찾는 연구도 진행되고 있다. 이 선행 연구는 단순히 각각의 증대 방법을 하나씩 적용하거나 모든 증대 방법을 섞어서 적용하는 것이 아니라 최적의 증대 방법의 조합을 찾아냄으로써 성능 향상에 이바지했다[2,3].

음향 처리 분야에서도 백색 잡음 추가, 재생 속도 변환, 그리고 소리의 피치 변환 방법 등 다양한 방법이 활용되고 있다[1,5]. 하지만 음향에서의 변형은 영상에 비해 많은 계산 복잡도를 지니고 있어서 영상 처리처럼 실시간 데이터 증대가 힘들다는 단점이 있다. 더욱이 일반적으로 심층 신경망에는 음향 데이터가 바로 들어가는 것이 아니라 푸리에 변환 및 필터링을 거친 스펙트로그램 (spectrogram)이 들어간다. 이 푸리에 변환에도 많은 시간이 소요돼서 학습 전에 미리 변환해놓기도 하는데, 이 역시 음향 데이터의 증대를 어렵게 만드는 요소이다. 최근엔 이미 변환된 스펙트로그램에서 일부분을 가리는 방법 (masking)을 적용하는 데이터 증대 방법이 연구되고 있다[6]. 영상 처리 분야와 유사하게 음성 인식에서도 최적의 데이터 증대 방법을 찾기 위한 시도가 있었지만, 앞서 언급한 음성에서 데이터 증대의 어려움 때문에 스펙트로그램에서 가능한 증대 방법만 적용했다는 아쉬운 점이 있다[7].

본 논문은 기존의 영상 및 음향에서 최적의 데이터 증대 방법 탐색에 영감을 받아 자연 소리 분류에서 최적의 데이터 증대 방법의 탐색을 연구한다. 특히, Hwang, Yeongtae, et al.[7]이 보여준 것처럼 음향 처리에서 데이터 증대를 통한 성능 변화는 영상 처리에서의 효과와 다르다는 것을 실험적으로 보여준다. 더 나아가서, 음향 데이터의 증대 영향을 고려하는 최적의 데이터 증대 방법을 실험적으로 찾는다. 결과적으로 자연 음향 데이터 세트 (ESC-50[8])에 최적화된 데이터 증대 방법을 적용함으로써, 자연 음향 분류 작업의 정확도를 향상하게 시킬 수 있었다.

Ⅱ. 본 론

1. 자연 음향 데이터 세트

본 논문의 학습에 사용된 자연 음향 데이터 세트 (ESC-50[8])는 실제 생활에서 마주할 수 있는 50가지의 다양한 객체 혹은 주변 상황의 음향을 포함하고 있다 (예를 들면, 기차, 비, 교회 종, 그리고 새 소리 등). 이 데이터 세트에는 총 2,000개의 녹음된 음향 데이터가 포함되어 있으며 각 음향은 44.1 kHz의 샘플링 레이트 (sampling rate)과 5초의 시간으로 일정하게 이루어져 있다. 전체 데이터 세트는 5개의 작은 세트 (fold)로 미리 구분되어 있어서, 학습 및 정확도를 산출할 때는 cross-validation을 통해 5번의 실험을 반복해서 평균한다. 데이터 세트 안의 1차원 음향 데이터는

a) 경희대학교 컴퓨터공학과(Computer Science and Engineering, Kyung Hee University)
‡ Corresponding Author : 배성호(Sung-Ho Bae)
E-mail: shbae@khu.ac.kr
Tel: +82-31-201-2593
ORCID: https://orcid.org/0000-0002-3389-1159
※ 이 논문의 연구 결과 중 일부는 “2020년 한국방송·미디어공학회 하계 학술대회”에서 발표한 바 있음.
※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2019-01-01768, Deep Neural Network based Real-Time Accurate Voice Source Localization using Drones).
· Manuscript received September 10, 2020; Revised November 16, 2020; Accepted November 16, 2020.

STFT (Short-Time Fourier Transform)에서 프레임은 40ms 단위이며 20ms가 겹치도록 프레임을 이동했다. 그리고 220개의 log-mel 필터 뱅크를 이용한 필터링을 통해 2차원의 스펙트로그램으로 변환되어 신경망의 학습에 입력으로 사용된다.

2. 합성곱 신경망 구조 및 학습 방법

학습에 사용한 심층 신경망의 구조는 그림 1과 같다. 처음엔 배치 정규화 (batch normalization)을 통해 데이터를 정제하고, 커널 크기 (kernel size)가 7인 합성곱 (convolution) 연산을 적용해 충분한 정보를 필터링하고 보존할 수 있도록 했다. 이어지는 배치 정규화과 ReLU 활성화 함수 계층을 통해 비선형성이 추가되고, Max Pooling 연산을 통해 중간 피쳐 (feature)의 크기 (resolution)가 줄어들게 된다. 4번 반복되는 구조에서는 커널 크기가 3인 합성곱이 2번 존재하는데 이는 커널 크기가 5인 합성곱 하나를 쓰는 것보다 계산 복잡도가 적으면서, 넓은 구간을 고려할 수 있기 때문이다. 중간의 drop-out 계층은 합성곱 신경망이 학습 과정에서 너무 학습 데이터에 치우치지 않고 일반화될 수 있게 도와주는 역할을 한다. 피쳐 추출의 마지막 부분에서는 커널 크기가 1인 합성곱을 통해 피쳐의 채널 수를 늘려서 global average pooling에서 발생할 수 있는 정보 손실을 최소화한다. 마지막의 완전 연결 계층 (fully connected layer)과 softmax 연산을 통해 최종 음향 분류 결과를 예측하게 된다.

본 논문에서 대부분의 기본 및 데이터 증대 실험은 2.1절의 ESC-50 데이터 세트와 2.2 절의 그림 1 합성곱 신경망 구조를 사용해서 진행했다. 신경망의 학습은 총 1600 epochs을 진행했으며 배치 크기는 64로 설정했다. 초기 러닝 레이트 (learning rate)은 0.001으로 설정했으며, 러닝 레이트는 학습 에폭 (epoch)에 따라 코사인 함수 형태로 감소한다[9]. 더욱 일반화된 합성곱 신경망을 학습시키기 위한 drop-out 계층은 확률적으로 중간 피쳐의 25% 값들을 매번 추론마다 임의로 다르게 제거한다. 모든 합성곱 계층에는 0.0001의 weight decay를, 완전 연결 계층에는 더욱 높은 0.1의 weight decay를 적용했으며, weight decay는 마찬가지로 모델의 일반화된 학습에 긍정적인 역할을 하게 된다. 서론에서 언급한 것처럼 음향 데이터는 실시간 증대가 어려운 점이 있으므로, 본 논문의 모든 데이터 증대 실험에서는 학습 전에 미리 기존 데이터 개수보다 6배 많게 증대해 놓고 학습을 진행한다.

3. 개별 데이터 증대 방법 실험

최적의 데이터 증대 방법을 찾기 위해 본 논문에서는 음향에 적용하는 증대 방법과 스펙트로그램에 적용하는 증대 방법을 모두 고려한다. 음향에 적용한 증대 방법은 백색 잡음 추가 (White Noise), 음향의 피치 변경 (Pitch Shift), 음향의 속도 변경 (Time Stretch), 그리고 음향의 시간 축 이동 (Time Shift)이 있다[1,5]. 스펙트로그램에 적용할 수 있는 증대 방법은 시간 가리기 (Time Mask)와 주파수 가리기 (Frequency Mask) 등이 있다[6,7]. 본 논문은 처음으로 두 가지 유형의 증대 방법을 모두 고려하며 영향을 관찰한다.

표 1은 위 문단에서 언급한 각각의 데이터 증대 방법을 실험한 결과를 보여준다. 첫 번째 열은 아무 데이터 증대 방법을 적용하지 않은 결과이며, 그 외의 열들은 각각의 증대 방법과 그에 따른 성능을 나타낸다. 표 1의 실험은 자연

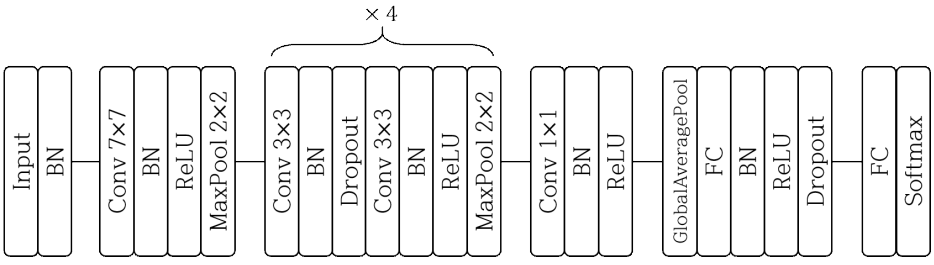

그림 1. 자연 음향 분류에 사용되는 합성곱 신경망의 구조도
Fig. 1. The architecture of convolutional neural network for environmental sound classification

음향 분류 작업에서 모든 데이터 증대 방법이 항상 분류

표 1. 자연 음향 데이터 세트에서 각각의 데이터 증대 방법에 대한 실험 결과
Table 1. The experimental results on ESC-50 dataset for each data augmentation method

Data Augmentation Method	Augmentation range	Top-1 Accuracy (%)	Difference
Baseline	-	85.95	-
White Noise	10 ~ 50 dB	85.2	-0.75
Pitch Shift	-1 ~ 1 step	87.15	1.2
Time Stretch	0 ~ 5 %	86.75	0.8
Time Shift	0 ~ 5 %	87.1	1.15
Time Mask	0 ~ 5 %	84.9	-1.05
Frequency Mask	0 ~ 10 bins	86.35	0.4

정확도의 향상을 불러오지 않는다는 것을 알 수 있다. 스펙트로그램에서 적용하는 데이터 증대 방법은 음향에 적용하는 방법보다 계산량이 적어서 선호될 수 있지만, 자연 음향 데이터 세트에서는 선행 연구[6]에서 음성 인식 작업에 적용했을 때만큼의 좋은 효과를 보여주지는 않았다. 이는 기존 이미지 데이터 증대 탐색 연구[2,3]처럼, 음향 처리 분야에서도 다른 종류의 작업마다 효과적인 데이터 증대 방법이 다를 수 있다는 것을 보여준다. 자연 음향 데이터 세트에서는 음향의 피치 변경 방법이 기준보다 1.2%포인트라는 가장 높은 성능 향상을 보여주었다. 다음 절에서는 각각 증대 방법의 효과가 아닌 여러 증대 방법을 동시에 적용했을 때의 효과를 탐색한다.

4. 복합 데이터 증대 방법 실험

이번 절에서는 여러가지의 데이터 증대 방법을 복합적으로 사용하며 실험을 진행했다. 복합적 사용을 위해서는 두 가지: i) 학습에 사용될 총 데이터 증대 방법 종류 ii) 학습 중 하나의 데이터에 동시에 적용될 데이터 증대 방법의 수가 결정되어야 한다. 먼저 학습에 사용되는 총 데이터 증대 방법 종류는 2.3 절에서 분류 성능 향상에 도움이 된 4 가지의 증대로 고정했다 (표 2의 세 번째 행의 첫 번째 열). 그 이유는 성능의 하락을 야기하는 개별 증대 방법은 복합적

표 2. 자연 음향 데이터 세트에서 다양한 복합적 데이터 증대 방법에 대한 실험 결과
Table 2. The experimental results on ESC-50 dataset for various combinations of data augmentation methods

Data Augmentation Method	The number of combinations	Top-1 Accuracy (%)	Difference
Baseline	-	85.95	-
Pitch Shift Time Stretch Time Shift Frequency Mask	1	89	3.05
	2	88.15	2.2
	3	88.35	2.4
	4	88.35	2.4

시도에서도 같은 경향을 보여주었기 때문이다. 추가로, 학습 중 하나의 데이터에 동시에 적용될 데이터 증대 방법의 수를 1부터 4까지 다양하게 바꾸며 실험을 진행했다. 예를 들어, 증대 방법을 2개로 하는 실험에서는, 하나의 음향 데이터를 증대하는데 있어서 총 4개의 증대 종류 중 무작위로 중복없이 2개를 선택해 순차적으로 적용한다[3]. 표 2는 복합적 증대 방법의 결과를 보여준다. 두 번째 행은 동시에 적용될 데이터 증대 방법의 수를 나타낸다.

복합적 증대의 모든 경우가 개별 증대보다 높은 성능 향상을 보여주는데, 이는 다양한 증대 방법이 복합적으로 활용될 경우 더욱 일반화된 신경망을 학습시키는 것이 가능하며 분류 정확도 향상에 기여가 크다는 것을 보여준다. 그런데 조합의 개수가 커서 많은 증대 방법이 동시에 사용될 경우, 오히려 성능 향상의 정도가 낮아지는 것을 볼 수 있다. 이는 심한 데이터의 변형은 오히려 성능 향상을 방해할 수 있다는 것을 보여준다. 결과적으로 동시에 1개의 증대 방법만 사용하여 복합적인 증대를 했을 때 89%의 정확도를 보여주었으며, 데이터 증대 방법을 적용하지 않은 합성곱 신경망 (85.95%)보다 정확도가 3.05%포인트 높아질 수 있었다.

5. 복잡한 신경망에서의 실험

2.4절은 다양한 데이터의 증대 방법에서의 효과에 집중하기 위해 그림 1과 같은 간단한 합성곱 신경망을 사용했었다. 이 2.5절에서는 그보다 복잡한 신경망에서도 복합적 데

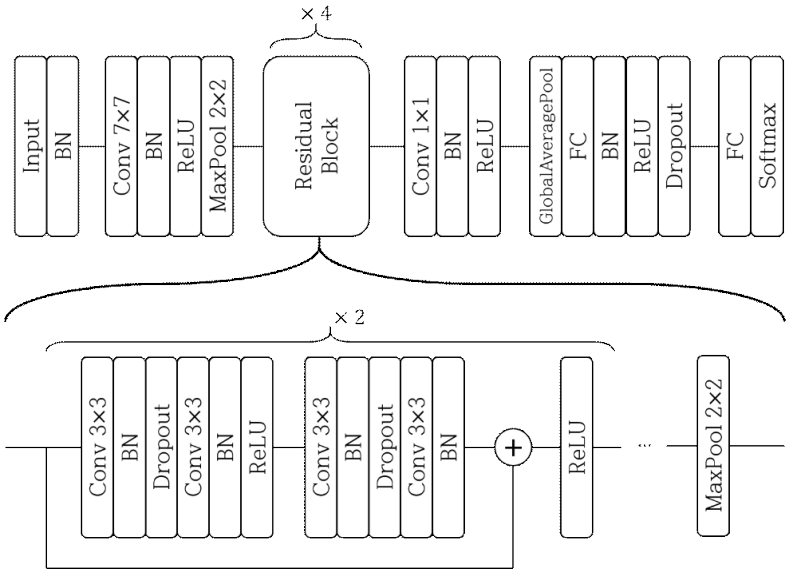

그림 2. 자연 음향 분류에 사용되는 복잡한 심층 신경망의 구조도
Fig. 2. The architecture of advanced convolutional neural network for environmental sound classification

이터 증대 방법이 효과가 있는지 검증하기 위한 실험을 진행했다. 실험에 사용된 복잡한 합성곱 신경망을 묘사하는 그림 2를 보면 기존 합성곱 신경망인 그림 1과 비교했을 때, 더욱 복잡하며 깊은 합성곱 연산 계층이 존재한다. 그림 1에서 네 번 반복되는 구조가 그림 2에서는 residual block으로 대체되며, 그 안의 두 계층의 결과 값을 더하는 연산은 합성곱 신경망을 깊게 쌓아도 원활한 학습이 가능하게 도와준다.

표 3은 2.4절에서 실행한 복합적 증대와 동일한 방법을 그림 2의 복잡한 합성곱 신경망에 적용한 결과를 기술한다. 이 실험은 복합적 데이터 증대 방법이 다른 형태의 신경망에서도 유효한지 알아보기 위한 목적으로 실시되었다. 또한, 일반적으로 깊은 신경망은 더욱 세분화된 분류를 가능하게 하는데, 이 때의 효과를 알아보기 위한 목적도 있다. 결과적으로, 복잡한 합성곱 신경망에서 복합적 데이터 증대를 적용한 결과의 경향은 기존 합성곱 신경망과는 다른 경향을 보여준다. 복잡한 합성곱 신경망에서는 하나의 데이터에 동시에 적용될 데이터 증대 방법의 수가 4일 때, 그림 1의 신경망보다 1.05%포인트 높은 가장 향상된 성능을 보여주었다. 이를 통해, 신경망의 구조 및 깊이에 따라 데이터 증대 방법의 효과가 달라진다는 것을 알 수 있다. 주목할 만한 부분은 기존 합성곱 신경망은 동시에 사용된 개수가 1일 때 성능이 가장 좋았지만, 복잡한 신경망에선 개수가 4일 때였다는 것이다. 이는 신경망의 크기에 따라 수용할 수 있는 한계가 다르며, 기존 신경망은 데이터 증대를 통한 많은 변형을 수용하지 못 했지만, 복잡한 신경망은 그 많은 변형을 수용하며 오히려 성능 향상에 도움이 되었다는 것을 암시한다.

표 3. 자연 음향 데이터 세트에서 다양한 복합적 데이터 증대 방법에 대한 실험 결과
Table 3. The experimental results on ESC-50 dataset for various combinations of data augmentation methods

Data Augmentation Method	The number of combinations	Top-1 Accuracy (%)	Difference
Baseline	-	85.95	-
Pitch Shift Time Stretch Time Shift Frequency Mask	1	88.95	3
	2	89.05	3.1
	3	89.20	3.25
	4	90.05	4.1

6. 실험 결과 비교

표 4는 자연 음향 분류 작업에서 선행 연구의 실험 결과와 본 논문의 실험 결과 간의 차이를 보여준다. 최근의 선행 연구는 크게 두 가지 방법을 통해 정확도의 향상을 끌어냈다. 예를 들면, log-mel spectrogram 뿐만 아니라 푸리에 변환 전의 1차원 음향 혹은 Mel-Frequency Cepstral Coefficients (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), the Constant Q-transform (CQT) 등의 다양한 피쳐를 추출하고 학습에 사용했다[5,15]. Multi-stream을 통한 더욱 복잡한 신경망 구조나 channel 혹은 temporal attention 구조를 통해 정확도를 올린 시도도 있다[5,14,15]. 하지만 본 논문의 2.4 절에서는 데이터 증대 효과에 집중하기 위해 단 하나의 log-mel feature extractor만 사용하며 multi-stream 혹은 attention 등의 복잡한 신경망 구조는 사용하지 않았었다. 그런데도 효과적인 데이터 증대만으로 자연 음향 데이터 세트에서 89%포인트의 정확도를 달성할 수 있었다. 이는 음향 분류 작업에서 최적화된 데이터의 변환 및 증대를 통해 더욱 일반화된 모델을 학습시키는 것이 얼마나 중요한 역할을 하는지 보여준다. 더 나아가서 2.5 절에서 복잡한 합성곱 신경망을 통한 실험은 기존 합성곱 신경망보다 1.05%포인트 높은 90.5%포인트의 정확도를 보여준다. 이 결과는 최적의 데이터 증대 탐색 방법이 다양한 크기의 신경망에도 보편적으로 적용될 수 있다는 것을 알려준다.

표 4. 자연 음향 데이터 세트에 대한 선행 연구와 본 논문의 학습 결과 비교
Table 4. Comparison of experimental results between prior work and our proposal on Esc-50 dataset

Method	Top-1 Accuracy (%)
Human[8]	81.3
AlexNet[10]	65
GoogleNet[10]	73
EnvNet2 + strong augment[11]	84.7
SoundNet[12]	74.2
CNN + Augment + Mixup[13]	83.9
CRNN + channel & temporal Attention[14]	86.5
Multi-stream + temporal Attention[15]	84
Multiple Feature + CNN with Attention[5]	88.5
Searched Augment + CNN (Ours)	89
Searched Augment + Residual CNN (Ours)	90.05

Ⅲ. 결 론

본 논문에서는 음향 처리 분류 작업에서 최적의 데이터 증대 방법을 탐색하기 위해 다양한 실험 진행하며, 그렇게 찾아진 방법이 얼마나 효과적인지 선행 연구와의 비교를 통해 보여준다. 탐색하는 과정은 처음에 다양한 데이터 증대 방법을 하나씩 적용해보고 그중에 정확도를 향상하게 시키는 방법들을 분별했다. 다음엔 이들을 복합적으로 적용하며 조합 개수를 다양하게 실험함으로써 어떤 증대 방법이 가장 효과적인지 탐색했다. 결과적으로, 최근 선행연구에서 제안한 multi-stream 혹은 attention 등의 고도화된 신경망 구조를 활용하거나 다양한 feature extractor를 통해 다양한 특성을 신경망에 제공해주는 것 없이 탐색한 데이터 증대 방법만으로도 자연 음향 데이터 세트 (ESC-50)의 분류 작업에서 89%포인트라는 높은 분류 정확도를 얻을 수 있었다. 더 나아가서 복잡하고 깊은 신경망을 사용했을 때는 복합적 데이터 증대를 통해 90%포인트의 정확도를 달성할 수 있었다. 이는 최적의 데이터 증대 방법이 작업 종류, 데이터 세트 및 신경망의 종류마다 달라질 수 있다는 것을 보여주며, 특히 음향 분류 작업에서 최적의 데이터 변형 및 증대가 신경망의 일반화 및 성능 향상에 얼마나 중요한 역할을 하는지 보여준다.

참 고 문 헌 (References)

[1] Salamon, Justin, and Juan Pablo Bello. "Deep convolutional neural networks and data augmentation for environmental sound classification." IEEE Signal Processing Letters, 24(3), pp.279-283, Jan 2017.

[2] Cubuk, Ekin D., et al. "Autoaugment: Learning augmentation strategies from data." Proceedings of the IEEE conference on computer vision and pattern recognition. May 24 2018.

[3] Cubuk, Ekin D., et al. "Randaugment: Practical automated data augmentation with a reduced search space." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702-703. 2020.

[4] Hendrycks, Dan, et al. "Augmix: A simple data processing method to improve robustness and uncertainty." arXiv preprint arXiv:1912.02781, Dec 5 2019.
[5] Sharma, Jivitesh, Ole-Christoffer Granmo, and Morten Goodwin. "Environment Sound Classification using Multiple Feature Channels and Deep Convolutional Neural Networks." arXiv preprint arXiv:1908.11219, Aug 28 2019.
[6] Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." arXiv preprint arXiv:1904.08779, Apr 18 2019.
[7] Hwang, Yeongtae, et al. "Mel-spectrogram augmentation for sequence to sequence voice conversion." arXiv preprint arXiv:2001.01401, Jan 6 2020.
[8] Piczak, Karol J. "ESC: Dataset for environmental sound classification." Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018, Oct 13 2015.
[9] I lya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, Aug 13 2016.
[10] Venkatesh Boddapati, Andrej Petef, Jim Rasmusson, and Lars Lundberg. Classifying environmental sounds using image recognition networks. Procedia Computer Science, 112:2048 – 2056, Jan 1 2017.
[11] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. CoRR, abs/1711.10282, 2017.
[12] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, pp. 892 – 900, 2016.
[13] Zhichao Zhang, Shugong Xu, Shan Cao, and Shunqing Zhang. Deep convolutional neural network with mixup for environmental sound classification. In Jian-Huang Lai, Cheng-Lin Liu, Xilin Chen, Jie Zhou, Tieniu Tan, Nanning Zheng, and Hongbin Zha, editors, Pattern Recognition and Computer Vision, pp. 356 – 367, 2018.
[14] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao. Learning attentive representations for environmental sound classification. IEEE Access, 7:130327 – 130339, 2019.
[15] Xinyu Li, Venkata Chebiyyam, and Katrin Kirchhoff. Multi-stream network with temporal attention for environmental sound classification. CoRR, abs/1901.08608, 2019.

저 자 소 개

박 진 배

- 2020년 3월 ~ 현재 : 경희대학교 컴퓨터공학과 석사과정
- 2013년 3월 ~ 2019년 2월 : 경희대학교 전자정보대학 컴퓨터공학 및 전자공학 공학사 (복수전공)
- ORCID : https://orcid.org/0000-0003-3469-199X
- 주관심분야 : 심층 신경망 양자화 및 가지치기

Teerath Kumar

- 2019년 3월 ~ 현재 : 경희대학교 컴퓨터공학과 석박통합과정
- ORCID : https://orcid.org/0000-0001-8769-4989
- 주관심분야 : 비지도 학습, 데이터 증대 및 생성

배 성 호

- 2017년 9월 ~ 현재 : 경희대학교 전자정보대학 컴퓨터공학과 조교수
- 2016년 7월 ~ 2017년 8월 : MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) 박사 후 연구원
- 2011년 2월 ~ 2016년 8월 : KAIST 전기 및 전자공학과 공학박사
- 2004년 3월 ~ 2011년 2월 : 경희대학교 전자정보대학 컴퓨터공학 및 전자공학 공학사 (복수전공)
- ORCID : https://orcid.org/0000-0002-3389-1159
- 주관심분야 : 심층 신경망 모델 압축/해석/탐색, 이미지 신호의 역문제