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

*박진배 **Teerath Kumar ***배성호

경희대학교

*qkrwlsqo94@gmail.com **teerathkumar142@gmail.com ***shbae@khu.ac.kr

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

*Jinbae Park **Teerath Kumar ***Sung-Ho Bae

Kyung Hee University

요약

심층 신경망은 영상 분류, 음성 인식, 그리고 문자 번역 등 다양한 분야에서 효과적인 성능을 보여주고 있다. 신경망의 구조 변화, 신경망 간의 정보 전달, 그리고 학습에 사용되는 데이터 증대 등의 확장된 연구를 통해 성능은 더욱 발전하고 있다. 그 중에서도 데이터 증대는 기존에 수집한 데이터의 변형을 통해 심층 신경망에 더 다양한 데이터를 제공함으로써 더욱 일반화된 신경망을 학습시기키는 것을 목표로 한다. 하지만 기존의 음향 관련 신경망 연구에서는 모델의 학습에 사용되는 데이터 증대 방법의 연구가 영상 처리 분야만큼 다양하게 이루어지지 않았다. 최근 영상 처리 분야의 데이터 증대 연구는 학습에 사용되는 데이터와 모델에 따라 최적의 데이터 증대 방법이 다르다는 것을 실험적으로 보여주었다. 이에 영감을 받아 본 논문은 자연에서 발생하는 음향을 분류하는데 있어서 최적의 데이터 증대 방법을 실험적으로 찾으며, 그 과정을 소개한다. 음향에 잡음 추가, 피치 변경 혹은 스펙트로그램의 일부 제한 등의 데이터 증대 방법을 다양하게 조합하는 실험을 통해 경험적으로 어떤 증대 방법이 효과적인지 탐색했다. 결과적으로 ESC-50 자연 음향 데이터셋에 최적화된 데이터 증대 방법을 적용함으로써 분류 정확도를 89%로 향상시킬 수 있었다.

1. 서론

심층 신경망의 학습에 있어서 데이터는 매우 중요한 역할을 한다. 훈련에 사용되는 데이터가 충분하지 않을 경우, 훈련된 신경망은 일반적이지 않은 훈련 데이터에 치우친 예측을 해서 실제의 모든 상황에서 알맞게 활용될 수 없게 된다[8]. 이를 보완하기 위해 데이터 증대 방법이 사용된다. 학습을 위해 수집한 데이터의 변형을 통해 더욱 다양한 형태로 만들어서 이를 통해 학습된 신경망이 더욱 일반적인 예측을 할 수 있도록 도와주는 기법이다[1, 2].

최근 객체 인식, 검출 및 분할화와 같은 영상 데이터를 입력으로 하는 심층 신경망을 활용하는 연구에서는 데이터 증대에 관련된 연구도 활발히 진행되고 있다[1, 2]. 영상의 좌우 반전, 회전, 혹은 명암 효과 등 다양한 변형을 시도할 뿐만 아니라 이들의 조합에 따른 영향도 연구되고 있다. 더 나아가서는 학습에 사용되는 데이터와 심층 신경망의 특성에 따라 최적의 데이터 증대 방법 및 조합이 다르다는 것을 실험을 통해서 증명하였다[1, 2]. 이들은 단순히 각각의 증대 방법을 하나씩 적용하거나 모든 증대 방법을 섞어서 적용하는 것이 아니라 최적의 증대 방법의 조합을 찾아냄으로써 성능 향상에 기여했다.

음향 처리에서도 백색 잡음 추가, 소리 속도 변환, 그리고 소리의 피치 변환 방법 등 다양한 방법이 활용된다[8]. 하지만 음향에서의 변형은 많은 계산 복잡도를 지닌다는 단점이 있다. 더욱이 일반적으로 심층 신경망에는 음향 데이터가 바로 들어가는 것이 아니라 푸리에 변환 및 필터링을 거친 스펙트로그램 (spectrogram)이 들어간다. 따라서, 이 스펙트로그램에서의 일부분을 가리는 방법

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2018R1C1B3008159)

본 논문은 과학기술정보통신부 및 정보통신산업진흥원의 '고성능 컴퓨팅 지원' 사업으로부터 지원받아 수행하였음

(masking) 등을 통한 데이터 증대는 최근에 연구되고 있다[6]. 음성 인식에서도 최적의 데이터 증대 방법을 찾기 위한 시도가 있었지만 스펙트로그램에서만 가능한 증대 방법을 적용했다는 아쉬운 점이 있다[3].

본 논문은 기존의 영상 및 음향에서 최적의 데이터 증대 방법 탐색에 영감을 받아 자연 소리 분류에 있어서 최적의 데이터 증대 방법의 탐색을 연구한다. 특히, Hwang, Yeongtae, et al. [3]이 보여준 것처럼 음향에서 데이터 증대의 영향을 영상에서의 증대와 다르다는 것을 실험적으로 보여준다. 그리고 음향 데이터의 증대 영향을 고려하는 최적의 데이터 증대 방법을 실험적으로 찾는다. 결과적으로 ESC-50 자연 음향 데이터셋에 최적화된 데이터 증대 방법을 적용함으로써 음향 분류 작업의 정확도를 89%로 향상시킬 수 있었다.

2. 방법 및 실험

2.1. 자연 소리 데이터

신경망 학습에 사용한 자연 음향 데이터셋 ESC-50 [7]은 실제 생활에서 마주할 수 있는 50가지의 다양한 객체 혹은 주변 상황의 음향을 포함하고 있다 (예를 들면, 기차, 비, 교회 종, 새 소리 등). 총 2000개의 음향 데이터가 있으며 각 음향은 5초로 일정하다. 전체 데이터는 5개의 fold로 미리 구분되어 있어서, 학습 및 결과를 산출할 때는 cross-validation을 통해 5번의 실험을 반복해서 평균한다[7]. 기존 1차원의 음향 데이터는 STFT(Short-Time Fourier Transform) 및 log-mel 필터링을 통해 2차원의 스펙트로그램으로 변환되어 신경망의 학습에 입력으로 사용된다.

2.2. 심층 신경망 구조

본 논문에서 사용한 심층 신경망의 구조는 그림 1과 같다. 처음엔 batch normalization을 통해 데이터를 정규화하고, 바로 7*7 convolution filter를 적용해 충분한 정보를 필터링하고 보존될 수 있도록 했다. 다음은 kernel size가 3인 convolution이 2번 이어지는데 이는 kernel size가 5인 convolution 하나를 쓰는 것보다 계산 복잡도가 줄면서 마찬가지로 넓은 구간은 고려할 수 있다. 중간의 drop-out 층은 신경망이 학습과정에서 너무 학습 데이터에 치우치지 않도록 도와주는 역할을 한다. Convolution 계산 후의 ReLU 활성화 함수는 비선형성을 부여하며, Max pooling 층은 중간 피쳐 (feature)의 크기를 줄여주는 역할을 한다. 이와 같은 구조가 4번 반복되면, point-wise convolution을 통해 channel 수를 늘리고 global average pooling을 진행한다. 마지막의 full-connect layer와 softmax 연산을 통해 최종 음향 분류 결과를 예측하게 된다.

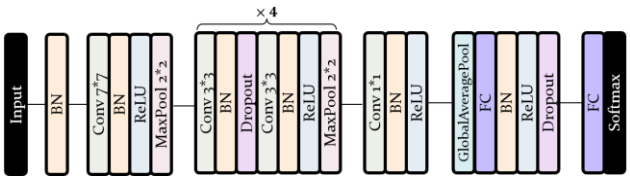

그림 1. 자연 음향 분류에 사용되는 심층 신경망의 구조도

2.3. 신경망 학습 방법

모든 기본 및 데이터 증대 실험은 2.1의 ESC-50 [7] 데이터셋과 2.2의 신경망 구조를 사용해서 진행된다. 신경망의 학습을 위한 상세 설명으로 실험은 1600 epoch을 진행했다. 초기 learning rate은 0.001으로 설정했으며, learning rate은 학습 epoch에 따라 cos 함수 형태로 변화한다[4]. 더욱 일반화된 학습을 위해 Drop-out 층은 25%의 값들을 매번 추론마다 임의로 다르게 제거한다. Convolution 층에는 0.0001의 weight decay를, fully-connected 층은 더욱 높은 0.1의 weight decay를 적용했다.

2.4. 각각의 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[8]. 스펙트로그램에 적용한 증대 방법은 시간 가리기 (Time Mask)와 주파수 가리기 (Frequency Mask)가 있다.[6] 본 논문은 처음으로 두 가지 유형의 증대 방법을 모두 고려하며 영향을 관찰한다.

표 1. 각각의 증대 방법의 실험 결과

데이터 증대 방법	증대 정도	정확도 (%)	차이
None	-	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

표 1은 위에서 언급한 각각의 데이터 증대 방법을 실험한 결과를 보여준다. 첫 번째 열의 결과는 아무 데이터 증대를 적용하지 않은 결과이며, 그 외의 열들은 각각의 증대 방법의 효과를 나타낸다. 표 1의 실험들은 모든 데이터 증대 방법이 항상 분류 정확도의 향상을 불러오지 않는다는 것을 알 수 있다.

2.5. 복합적 증대 방법 실험

더 나아가서 여러 증대 방법을 복합적으로 얼마나 많이 동시에 적용했을 때 및 어떤 순서로 적용했을 때의 영향을 관찰하고 어떻게 조합해야 더욱 최적의 증대가 가능한지 알기 위한 실험을 진행했다. 표 2는 2-3의 실험에서 성능 향상에 긍정적이었던 증대 방법만 선택하고, 여러 증대를 동시에 복합적으로 적용했을 때의 실험 결과를 보여준다. 예를 들어, 조합 개수가 2인 경우, 선택된 4개의 증대 방법들 중 매번 2개를 무작위로 선정해서 데이터 증대를 진행하게 된다[2]. 표 2의 결과를 보면, 여러 증대 방법들이 복합적으로 사용될 경우, 표 1처럼 하나의 방법만 사용하는 것보다 신경망을 훨씬 일반적으로 학습시킬 수 있다는 것을 보여준다. 하지만, 조합 개수가 커서 많은 증대 방법이 동시에 사용될 경우, 오히려 데이터의 변형이 심해져 성능 향상의 정도가 낮아지는 것을 볼 수 있다. 결과적으로 동시에 1개의 증대 방법만 사용하여 복합적인 증대를 했을 때, 가장 높은 효과를 볼 수 있었다.

표 2. 다양한 증대 방법의 복합적 적용 실험 결과

증대 방법 집합	조합 개수	정확도 (%)	차이
None	-	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

2.6. 실험 비교

표 3은 선행 연구의 실험 결과와 본 논문의 실험 간의 비교를 보여준다. 많은 선행 연구들은 높은 정확도를 달성하기 위해 다양한 feature extraction을 하거나 channel or temporal attention을 적용한다[8, 10, 13, 14]. 하지만 본 논문에서는 데이터 증대 효과에 집중하기 위해 단 하나의 log-mel feature extractor만 사용하며 attention 혹은 multi-stream 등의 복잡한 신경망 구조는 사용하지 않았다. 그럼에도 불구하고 효과적인 데이터 증대만으로 선행 연구보다 뛰어난 정확도를 가지는 분류 모델을 학습시킬 수 있었다. 이는 음향 분류 작업에서 데이터의 변환 및 증대를 통한 일반화된 모델을 학습시키는 것이 얼마나 중요한 것인지 보여준다.

표 3. 선행 연구와의 성능 비교

모델	정확도 (%)
Human [5]	81.3
AlexNet [9]	65
GoogleNet [9]	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 [10]	84
Multiple Feature + CNN with Attention [8]	88.5
CNN + Searched Augment (Ours)	**89**

3. 결론

본 논문에서는 음향 처리에서 최적의 데이터 증대 방법을 탐색해보고 그 증대 방법의 적용이 효과적임을 다양한 실험을 통해 보여준다. 처음엔 다양한 데이터 증대 방법을 각각 적용해보고 그 중에 긍정적인 방법들을 분류했다. 다음엔 이들을 복합적으로 적용하며 조합 개수를 다양하게 실험했을 때, 어떤 증대 방법이 가장 효과적인지 탐색했다. 결과적으로, multi-stream 혹은 attention 등의 고도화된 신경망 구조 활용이나 다양한 feature extractor를 통한 다양한 특성 제공없이 데이터 증대만으로 자연 음향 분류 작업 (ESC-50, [7])에서 89%라는 가장 좋은 성능을 얻을 수 있었다. 이는 음향 데이터에 있어서 데이터 변형 및 증대를 통한 일반화가 성능 향상에 얼마나 중요한 역할을 하는지 보여준다.

참고 문헌

[1] Cubuk, Ekin D., et al. "Autoaugment: Learning augmentation strategies from data." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019.

[2] Cubuk, Ekin D., et al. "RandAugment: Practical data augmentation with no separate search." *arXiv preprint arXiv:1909.13719* (2019).

[3] Hwang, Yeongtae, et al. "Mel-spectrogram augmentation for sequence to sequence voice conversion." *arXiv preprint arXiv:2001.01401* (2020).

[4] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

[5] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd Annual ACM Conference on Multimedia, pages 1015–1018. ACM Press.

[6] Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." *arXiv preprint arXiv:1904.08779* (2019).

[7] Piczak, Karol J. "ESC: Dataset for environmental sound classification." *Proceedings of the 23rd ACM international conference on Multimedia*. 2015.

[8] 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* (2019).

[9] Venkatesh Boddapati, Andrej Petef, Jim Rasmusson, and Lars Lundberg. Classifying environmental sounds using image recognition networks. Procedia Computer Science, 112:2048 – 2056, 2017. Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

[10] Xinyu Li, Venkata Chebiyyam, and Katrin Kirchhoff. Multi-stream network with temporal attention for environmental sound classification. CoRR, abs/1901.08608, 2019.

[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, pages 892–900, USA, 2016. Curran Associates Inc.

[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, pages 356–367, Cham, 2018. Springer International Publishing.

[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.