한국음향학회지 제39권 제5호 pp. 414~423 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.5 (2020)
https://doi.org/10.7776/ASK.2020.39.5.414

pISSN : 1225-4428
eISSN : 2287-3775

약한 레이블을 이용한 확장 합성곱 신경망과 게이트 선형 유닛 기반 음향 이벤트 검출 및 태깅 알고리즘

Dilated convolution and gated linear unit based sound event detection and tagging algorithm using weak label

박충호,[1] 김동현,[1] 고한석[1†]

(Chungho Park,[1] Donghyun Kim,[1] and Hanseok Ko[1†])

[1]고려대학교 전기전자공학과 지능신호처리 연구실

(Received July 24, 2020; accepted August 27, 2020)

초 록: 본 논문은 약한 레이블 기반 음향 이벤트 검출을 위한 시간-주파수 영역분할 맵 추출 모델에서 발생하는 희소성 및 수용영역 부족에 관한 문제를 완화 시키기 위해, 확장 게이트 선형 유닛(Dilated Convolution Gated Linear Unit, DCGLU)을 제안한다. 딥러닝 분야에서 음향 이벤트 검출을 위한 영역분할 맵 추출 기반 방법은 잡음 환경에서 좋은 성능을 보여준다. 하지만, 이 방법은 영역분할 맵을 추출하기 위해 특징 맵의 크기를 유지해야 하므로 풀링 연산 없이 모델을 구성하게 된다. 이로 인해 이 방법은 희소성과 수용영역의 부족으로 성능 저하를 보이게 된다. 이런 문제를 완화하기 위해, 본 논문에서는 정보의 흐름을 제어할 수 있는 게이트 선형 유닛과 추가의 파라미터 없이 수용영역을 넓혀 줄 수 있는 확장 합성곱 신경망을 적용하였다. 실험을 위해 사용된 데이터는 URBAN-SED와 자체 제작한 조류 울음소리 데이터이며, 제안하는 DCGLU 모델이 기존 베이스라인 논문들보다 더 좋을 성능을 보였다. 특히, DCGLU 모델이 자연 소리가 섞인 환경인 세 개의 Signal to Noise Ratio(SNR)(20 dB, 10 dB, 0 dB)에서 강인하다는 것을 확인하였다.
핵심용어: 음향 태깅, 음향 이벤트 검출, 확장 합성곱 신경망, 게이트 선형 유닛, 시간-주파수 영역분할 맵, 약한 레이블

ABSTRACT: In this paper, we propose a Dilated Convolution Gate Linear Unit (DCGLU) to mitigate the lack of sparsity and small receptive field problems caused by the segmentation map extraction process in sound event detection with weak labels. In the advent of deep learning framework, segmentation map extraction approaches have shown improved performance in noisy environments. However, these methods are forced to maintain the size of the feature map to extract the segmentation map as the model would be constructed without a pooling operation. As a result, the performance of these methods is deteriorated with a lack of sparsity and a small receptive field. To mitigate these problems, we utilize GLU to control the flow of information and Dilated Convolutional Neural Networks (DCNNs) to increase the receptive field without additional learning parameters. For the performance evaluation, we employ a URBAN-SED and self-organized bird sound dataset. The relevant experiments show that our proposed DCGLU model outperforms over other baselines. In particular, our method is shown to exhibit robustness against nature sound noises with three Signal to Noise Ratio (SNR) levels (20 dB, 10 dB and 0 dB).
Keywords: Audio tagging, Sound event detection, Dilated convolution, Gated linear unit, T-f segmentation map, Weak label

PACS numbers: 43.60.Bf, 43.64.Gf

†**Corresponding author:** Hanseok Ko (hsko@korea.ac.kr)
Engineering Building Room 419, Department of Electronics and Computer Engineering, Korea University Anam Campus, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea
(Tel: 82-2-3290-3239, Fax: 82-2-3291-2450)

I. 서 론

최근 인공지능 기술의 발전과 함께 음향 관련 연구들이 다수 진행되고 있다. 그중 음향 이벤트 인식 및 검출은 감시 시스템,[1] 사람과 컴퓨터 간의 상호작용[2] 및 야생 동물 모니터링[3-7]과 같이 국내외에서 다양한 분야에서 사용되고 있다. 이처럼 음향 이벤트 검출 문제는 다양한 인터페이스를 통해 폭넓게 적용될 수 있으므로 인공지능 분야에서 중요한 이슈로 대두되고 있다.

음향 이벤트 검출은 오디오에 어떤 소리 이벤트가 언제 발생했는지에 대한 답을 찾는 문제이다. 연구 초기에는 프레임별 이벤트 분류 문제를 기반한 딥러닝 학습 모델[8-10]들이 많이 제안되었다. 하지만 위의 모델들은 프레임 단위의 이벤트 발생 여부에 대한 정보(강한 레이블)가 필요하므로 대량의 데이터를 구성하는데 어려움이 있었다. 이러한 한계점을 완화 시키기 위해 약한 레이블(weak label) 기반 연구들이 진행되고 있다.[11-14] 약한 레이블은 오디오 클립에 어떤 이벤트가 포함되어 있는지에 대한 정보만을 바탕으로 기존의 강한 레이블을 이용한 프레임 레벨 분류가 아닌 인스턴스 레벨 멀티-레이블 분류를 수행한다.

약한 레이블 기반 음향 이벤트 검출기의 구조는 크게 이벤트 분류기와 소리 내의 이벤트의 시점과 종점을 검출하는 검출기로 구성된다. 이벤트 분류기는 일반적인 분류기에 사용되는 원-핫 레이블이 아닌 하나의 오디오 클립에 복수의 레이블을 가지는 멀티-레이블 분류기를 사용하며, 분류기의 저차원 특징 맵을 고차원으로 확장하여 소리의 발생 시점을 검출한다.[11] 그러나, 합성곱 신경망(Convolutional Neural Network, CNN)을 이용한 특징 추출 과정에서 시간 축에 대한 정보 및 특성이 없어지게 되므로, 위의 문제를 완화 시키기 위해 주파수 축으로만 차원을 줄이고, 시간 축에 대한 정보는 유지하는 학습법이 연구되었다.[12] 또한, 합성곱 신경망의 특징 추출 과정에 게이트 선형 유닛(Gated Linear Unit, GLU)[13]를 적용하여 레이어 별 활성화 함수를 학습한 구조가 향상된 성능을 보였다.[14] 하지만, 주파수 축 차원 축소 방법은 잡음에 민감하게 동작하여 잡음요소가 큰 환경에서 높은 성능 하락을 보인다. 이러한 문제를 줄이기 위해, 시간-주파수의 차원을 유지하는 학습법이 고안되었다.[14] 이 학습 방법은 입력 오디오 클립에서 목표가 되는 이벤트를 분리해 내는 영역분할 맵(segmentation map)을 찾아내는 방식으로, 특징 맵의 시간-주파수 차원을 유지하며 추출한 특징이 각 이벤트에 미치는 영향을 알 수 있도록 영역분할 맵을 클래스 개수만큼 추출한다. 모델의 마지막 레이어에서는 추출한 시간-주파수 영역분할 맵에 글로벌 풀링 연산을 적용하여 타겟 이벤트가 오디오 안에 있을 확률을 계산하게 된다. 이러한 학습 과정은 오디오 클립에서 타겟 이벤트에 해당하는 특징을 추출하기 때문에 잡음 환경에 강인한 성능을 보여준다.

위와 같은 이점이 있는 시간-주파수 차원을 보존하는 학습 방법이 잡음 환경에서 향상된 성능을 보여주지만, 영역분할 맵 추출 과정에서 다음과 같은 한계점이 있다

1. 시스템의 희소성 부족: 합성곱 신경망은 작은 크기의 커널과 특징 맵의 차원을 줄이는 풀링을 활용하여 패치 별 중요한 특징을 찾아가는 방식으로 학습한다. 하지만 Reference [14]는 풀링 없이 특징 맵의 차원을 유지하면서 학습하기 때문에, 시스템의 희소성이 줄어들게 된다.
2. 수용 공간 부족: 수용 공간은 합성곱 신경망 모델이 볼 수 있는 정보의 양을 나타낸다. 컴퓨터 비전 분야에서 많이 사용되는 모델들은[15,16] 특징 맵의 크기가 레이어를 지나면서 줄어들기 때문에 작은 크기의 커널을 통해 다양한 패턴을 학습할 수 있다. 그에 반해 Reference [14] 모델은 입력과 같은 크기의 영역분할 맵 추출을 위해 특징 맵 크기를 유지하면서 [3 × 3] 크기의 커널을 모든 레이어에서 동일하게 사용하기 때문에, 학습의 효율성이 떨어진다.

또한, Reference [14] 모델의 성능은 영역분할 맵에 의존적이므로 시스템 희소성 부족과 수용 공간 부족 문제는 시스템의 큰 성능 저하를 발생시킬 수 있다. 따라서 본 논문은 합성곱 신경망에 게이트 선형 유닛과 확장 합성곱 신경망을 활용해 위 문제에서 발생하는 성능 저하를 줄이고자 한다.

제안하는 모델에서는 기존 모델의 희소성에 대한 문제를 해결하기 위해, 게이트 선형 유닛[17]을 사용하였다. 게이트 선형 유닛은 합성곱 신경망의 커널을 통해 정보에 해당하는 특징과 추출된 특징이 유효한 정보인지 아닌지를 판단하는 게이트를 생성한다. 생성된 게이트는 시그모이드 함수를 통해 0과 1 사이 값을 가지게 되며, 생성된 특징과 요소별 곱셈을 통해 중요한 정보는 살리고 잡음에 해당하는 배경 소리를 차단하게 함으로써 정보의 흐름을 제어할 수 있다.

제한된 수용영역으로 인한 단일화 되는 패턴 학습에 대한 해결책으로는, 확장 합성곱 신경망[18]을 이용하였다. 확장 합성곱 신경망은 특징 맵의 크기를 유지하면서, Reference [14] 모델에 사용된 파라미터와 동일한 개수의 파라미터를 이용하여 넓고 다양한 수용영역을 구성할 수 있으므로, 제안하는 모델에 적용하였다.

또한, Reference [14]에서 실험에 사용한 데이터베이스는 실제 환경을 고려하지 않은, 이벤트가 겹치지 않는 상황만을 재연한 데이터를 사용하였다. 이에 본 논문에서는 이벤트가 같은 시점에 중복적으로 발생하는 데이터인 URBAN-SED 데이터[19]와 자체 제작한 조류 울음소리 데이터를 이용하여 실험하였다.

이후 논문은 II에서 제안하는 방법에 사용된 기술과 모델에 대한 설명을 기재하였고, III에서 음향 태깅, 음향 이벤트 검출, Signal-to-Noise Ratio(SNR)에 따른 성능을 비교한 토의 및 결과를 기재하였다. IV는 결론으로 마무리 지었다.

II. 제안하는 방법

2.1 약한 레이블

기존의 음향 이벤트 검출 알고리즘은 Fig. 1의 강한 레이블과 같이 오디오 클립 내에 이벤트의 발생 시점과 종점에 대한 정보가 포함된 데이터를 기반으로 학습을 하였다. 하지만 강한 레이블은 사람이 반복적으로 오디오 파일을 들으면서 이벤트 발생 여부와 정확한 발생 시간을 확인해야 하므로 강한 레이블 데이터를 만들기 위해서는 많은 시간과 노동이 필요하다. 그에 반해서 약한 레이블은 오디오 클립

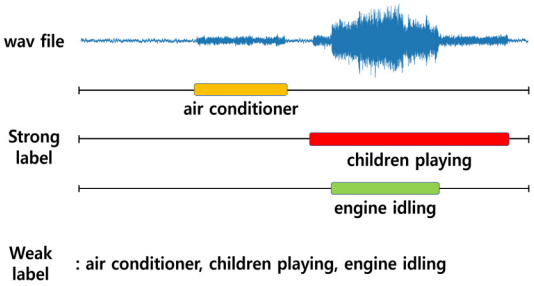

Fig. 1. (Color available online) This is an example of weak and strong labels. Strong labels include on–off sets for each class, and weak labels only contain information about whether an event has occurred.

내의 이벤트의 존재 여부에 대한 정보만 요구되기 때문에 강한 레이블보다 약한 레이블이 수월하게 데이터 작업이 가능하다. 이러한 이점 때문에, 다량의 데이터를 필요로 하는 딥러닝 분야에서 약한 레이블을 기반한 연구들이 늘어나고 있는 추세다.[11-14]

2.2 확장 합성곱 신경망

일반적인 합성곱 신경망은 커널의 파라미터 개수를 늘리지 않고 수용영역을 넓히기 위해 풀링 연산과 함께 사용된다. 하지만, 풀링 연산을 거쳐서 나온 특징은 특정 정보 이외의 정보는 손실되게 된다. 그에 반해, 확장 합성곱 신경망[18]은 풀링 연산 없이 Fig. 2와 같이 빨간 점에 해당하는 픽셀의 정보만을 이용하여 합성곱 신경망 연산을 수행하므로 입력 데이터에 대한 정보 손실 없이 수용영역을 넓혀 줄 수 있다. 본 논문에서는 이벤트별 시간 -주파수 도메인 영역 분할 맵을 추출하는 부분에 확장 합성곱 신경망을 사용하였다. 확장률은 Fig. 2에서 나타내었듯이 빨간 점들간의 거리를 의미하기 때문에, 높을수록 더 큰 수용영역을 가지게 된다. 제안하는 모델에서는 확장률을 1에서 최대 8까지 점차적으로 늘여가며 사용하였다.

2.3 게이트 선형 유닛

게이트 선형 유닛[13,17]은 합성곱 신경망을 통해 나온 특징에 대한 정보를 제어하는 게이트 역할을 함으로써, 시간-주파수 도메인의 정보 중 유의미한 정보에 해당하는 이벤트 정보는 전달하고, 무의미한 정보에 해당하는 배경 소리는 버림으로써 잡음에 보

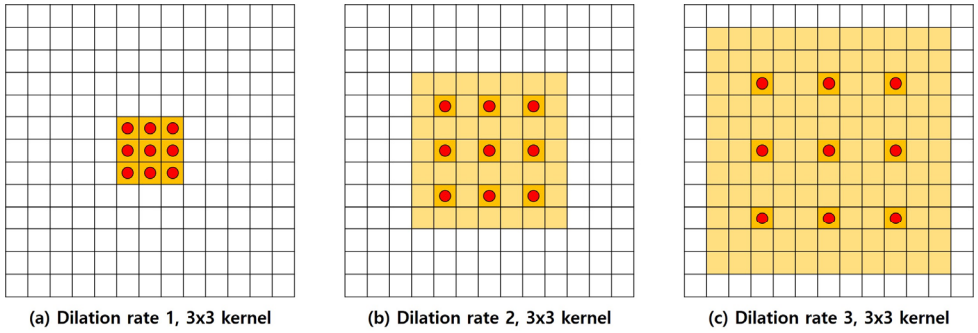

Fig. 2. (Color available online) This is an example of the kernel of dilated convolution. The kernel size is 3 x 3, and the size of the receptive field changes as the dilation rate changes.

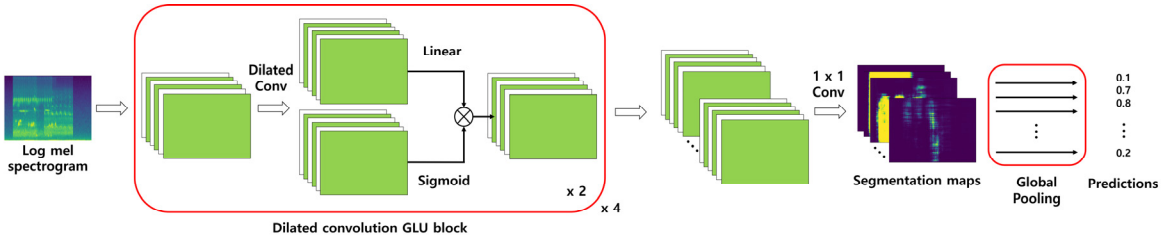

Fig. 3. (Color available online) The architecture of the proposed DCGLU model. This model uses log-Mel spectrogram as input data and extracts a segmentation map through 4 DCGLU convolution blocks and 1 × 1 convolution. The extracted segmentation map outputs the predictions through global pooling.

다 강인한 모델을 구성 하는데 도움을 줄 수 있다. 게이트 선형 유닛 수식으로 표현하면 Eq. (1)과 같다.

$$Y=(W_1*X+b_1)\odot\sigma(W_G*X+b_G), \quad (1)$$

σ는 시그모이드 함수이며, W_1와 W_G는 각각 정보와 게이트를 출력하는 합성곱 신경망의 커널을 나타내고 b_1와 b_G는 바이어스를 의미한다. X는 이전 레이어에서 출력된 특징이거나 입력되어 들어온 데이터이고 $\odot$는 요소별 곱셈이다. 즉, 게이트 선형 유닛은 합성곱 신경망에서 커널 개수를 두 배로 늘려 두 배의 특징을 추출하여, 특징의 절반을 정보로 이용하고, 나머지 절반은 시그모이드 함수를 거친 게이트로 사용함으로써 정보를 흐름을 제어할 수 있는 학습 가능한 활성화 함수로 기존의 활성화 함수를 대체되어 사용된다.

2.4 Dilated Convolution Gated Linear Unit 모델

본 논문에서 제안하는 모델은 Reference [14]의 영역분할 맵 추출 모델을 기반하여 앞서 소개한 약한 레이블을 이용한 확장 합성곱 신경망과 게이트 선형 유닛을 적용한 확장 게이트 선형 유닛 Dilated Convolution Gated Linear Unit(DCGLU) 모델이다. 모델은 구조는 Fig. 3과 같다. 입력 데이터인 로그 멜 스펙트로그램은 2개의 확장 합성곱 신경망과 활성화 함수를 게이트 선형 유닛으로 대체한 확장 게이트 선형 유닛 블록을 4번 반복하여 특징을 추출한다. 이때 확장 합성곱 신경망의 커널 크기는 3×3으로 하였고, 확장률은 첫 번째 블록에서 1, 두 번째 블록은 2, 세 번째 블록은 4, 네 번째 블록은 8로 하여 수행하였다. 각 레이어의 커널 개수는 실험을 통해 최적화하여 구성하였다. 자세한 모델 구성은 Table 1에 기재하였다. 확장 게이트 선형 유닛 블록을 통해 추출된 특징은 1 × 1 합성곱 신경망과 시그모이드 활성화 함수를 통해 클래스 개수만큼의 0~1 사이의 값을 가지는 영역분할 맵을 얻게 된다. 추출한 영역분할 맵은 Fig. 4와 같이 출력된다. 이렇게 얻은 영역분할 맵은 약한 레이블과의 이진 크로스 엔크로피 손실을 계산하기 위해 글로벌 풀링을 이용하여 클래스별 예측 값을 도출한다. 이때 글로벌 풀링은 글로벌 가중치 정렬

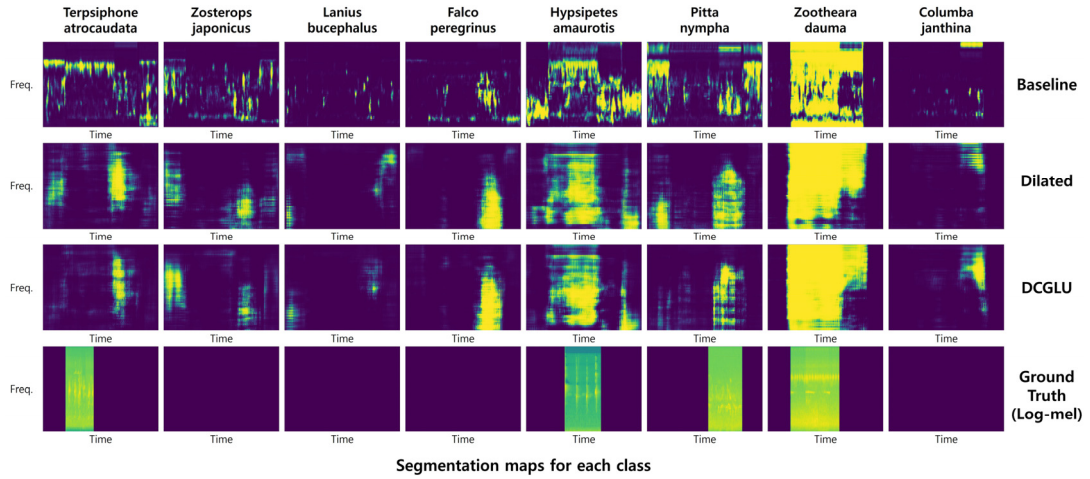

Fig. 4. (Color available online) This is the segmentation map of the bird sound data with SNR of 20, the biggest difference from the baseline in the experimental results. Ground truth is a log-Mel spectrogram for each class.

Table 1. Configuration of Dilated GLU CNN.

Layers {kernel size, dilation rate, repeat}	Number of kernel	Output Size {chan.× time × freq}
Input log mel spectrogram	-	1 × 431 × 64
Dilated CNN GLU {3 × 3, 1, 2}	64	32 × 431 × 64
Dilated CNN GLU {3 × 3, 2, 2}	128	64 × 431 × 64
Dilated CNN GLU {3 × 3, 4, 2}	256	128 × 431 × 64
Dilated CNN GLU {3 × 3, 8, 2}	256	128 × 431 × 64
CNN {1 × 1}, Sigmoid	10	10 × 431 × 64
Global Weighted Rank Pooling	-	10

풀링(Global Weighted Rank Pooling, GWRP)[14,20]을 사용하였다. 글로벌 가중치 정렬 풀링은 시간-주파수 영역분할 맵 추정에서 글로벌 최대 풀링(Global Max Pooling, GMP)와 글로벌 평균 풀링(Global Average Pooling, GAP)의 문제점인 과소 추정[20]과 과대 추정[20]에 대한 해결책으로, 추출된 특징의 각 픽셀 정보의 크기에 따라 가중치를 주어 풀링 연산에 반영되게 된다. 즉, 큰 값을 가지는 픽셀의 정보에 대한 가중치가 크고, 작은 값을 가지는 픽셀의 정보에 대한 가중치는 작게 함으로써 전체 픽셀에 대한 정보를 적절히 반영하여 풀링 연산을 하게 된다. 글로벌 가중치 정렬 풀링에 대한 수식은 Eq. (2)와 같다.

$$GWRP(S_k) = \frac{1}{Z(r)} \sum_{j=1}^{M} r^{j-1} (S_k)_j, \tag{2}$$

S는 1×1 합성곱 신경망을 거쳐서 나온 영역분할 맵을 의미하며, k는 클래스의 개수이다. j는 출력된 영역분할 맵의 각 픽셀값의 크기에 따라 높은 값 부터 낮은 값으로 내림차순 정렬한 인덱스이다. M은 시간 × 주파수로 전체 픽셀 개수를 나타낸다. r은 하이퍼 파라미터로 $0 \leq r \leq 1$의 값으로 설정하고, $Z(r) = \sum_{j=1}^{M} r^{j-1}$는 정규화를 위한 항이다. 위의 수식에 따라 영역분할 맵이 풀링 연산을 거치게 되면 0과 1 사이의 클래스 레벨 예측값을 출력하게 되고, 예측된 값은 약한 라벨과의 이전 크로스 엔트로피 손실을 계산하여 모델을 학습하게 된다.

III. 실 험

3.1 데이터베이스

실험에 사용한 데이터는 URBAN-SED[19]와 자체 제작한 조류 울음소리이다. URBAN-SED는 도시환경에서 발생할 수 있는 에어컨, 차 엔진, 드릴, 총소리 등 10가지 이벤트로 구성되어 있고, 오디오 클립의 구성은 배경 소리(잡음)가 브라운 노이즈이며, 오디오 클립 내에 1개~9개의 이벤트가 겹쳐서 발생하는 상황으로 되어있다. 각각의 오디오 클립의 길이는 10 s이다. 데이터 개수는 train, validation, test set을 포함하여 10,000개의 데이터로 약 30 h의 오디오 클립으로 구성된다.

조류 울음소리 데이터는 자체 수집하여 제작한 데이터로 8종의 클래스를 가지며, URBAN-SED 데이터의 상황과 같은 상황인 10 s 내에 1개 ~ 7개의 이벤트가 겹쳐서 발생하도록 하였다. 추가적으로, SNR에 따른 비교 실험을 위해 조류 울음소리 데이터의 test set에 실제 자연환경에서 발생할 수 있는 잡음인 폭우, 비, 바람, 강 소리를 섞은 데이터를 생성하여 실험하였다. 잡음 데이터의 SNR은 0, 10, 20으로 하여 잡음 상황 성능을 측정하였다.

3.2 특징 추출과 하이퍼 파라미터

오디오 클립의 샘플링 레이트는 44,100 Hz이며, 윈도우 크기 2048, 홉 크기 1024로 하여 Short-Time Fourier Transform(STFT)를 통해 스펙트로그램을 생성하였다. 위의 같은 구성이 시간과 주파수 도메인에서 좋은 해상도를 가진다고 알려져 있다.[21] 생성한 스펙트로그램은 64개의 멜 필터뱅크를 이용하여 멜 스펙트로그램을 추출 후, 로그를 씌워 음향 관련 연구에서 많이 사용되고 있는 로그 멜 스펙트로그램(log Mel spectrogram)[22,23]을 입력 데이터로 사용하였다.

모델을 학습할 때, batch size는 8로 하였고, 학습률은 0.001이며, Adam Optimizer[24]를 사용하였다. 또한, 모델의 안정화와 학습 속도 향상을 위해 1 × 1 합성곱 신경망 레이어를 제외한 모든 레이어에서 batch normalization[25]을 적용하였고, 글로벌 가중치 정렬 풀링(GWRP)에서 r는 0.9998로 하여 학습하였다.

3.3 평가 지표

평가 지표로는 F1 score, Area Under the Curve(AUC), mean Average Precision(mAP)을 이용하여 음향 태깅, 음향 이벤트 검출에 대한 성능을 검증하였다. F1 score는 시스템의 관점에서의 정확도를 나타내는 정밀도와 데이터 관점에서의 정확도를 나타내는 재현율을 기반하여 계산된다. F1 score는 정밀도와 재현율이 둘 중 하나라도 낮은 값을 가지게 되면 0에 가까운 값을 도출하며, 둘 다 높은 수치를 가져야만 1에 가까운 값을 출력한다.

AUC는 참 양성 비율과 허위 양성 비율을 그래프로 그린 Receiver Operating Characteristic(ROC) 커브[26]의 아래 영역의 넓이로 하나의 수치로 나타내어진다. AUC는 사용하는 경우 수동으로 기준치를 지정할 필요가 없으며, 시스템이 학습되지 않은 경우와 같이 랜덤한 값을 출력할 경우 0.5의 수치를 가지게 된다.

AP는 AUC와 마찬가지로 정밀도와 재현율에 대한 그래프의 아래 영역의 넓이를 의미한다. F1 score와 마찬가지로 정밀도와 재현율이 높을수록 1에 가까운 값을 가지며, 단일 이벤트 검출이 아닌 다중 이벤트 검출의 경우 각 클래스의 AP 평균(mAP)으로 성능을 평가한다.[27]

3.4 음향 태깅 및 이벤트 검출 성능 평가

음향 태깅에 대한 결과는 모델을 통해 추출된 시간-주파수 영역분할 맵에서 글로벌 가중치 정렬 풀링을 통해 출력된 각 이벤트 클래스가 오디오 클립내에 있을 확률이 0.1 이상일 때 이벤트가 있다고 판단을 하고 0.1 미만의 경우 이벤트가 없다고 판단하여 성능을 측정한다. 수식으로 표현하면 Eq. (3)과 같다.

$$e_k = \begin{cases} 1 & \text{if} \quad GWRP_k > 0.1 \\ 0 & otherwise \end{cases}. \tag{3}$$

$GWRP$는 Eq. (2)로 계산된 이벤트 클래스에 대한 확률이며, k는 각 이벤트 클래스에 해당하는 인덱스이다.

시간별 음향 이벤트 검출 결과는 시간-주파수 도메인에서 각각의 이벤트에 대한 정보를 담고 있는 추출된 시간-주파수 영역분할 맵에서 각 시간 t에 포함된 주파수축에 대한 정보들의 합의 평균이 0.1 이상이면 이벤트가 존재한다고 여기게 된다. 또한, 음향 이벤트 검출 결과는 음향 태깅 결과에서 이벤트가 있다고 판단되는 클래스의 경우만을 고려하여 측정하였다. 시간 t에 대한 이벤트 발생 여부 판단을 위한 수식은 Eq. (4)와 같다.

$$v_k(t) = \frac{1}{N}\sum_{i=1}^{N} S_k(t, f_i). \tag{4}$$

N는 주파수 축의 크기, t는 시간을 의미하며, S는 영

역분할 맵, k는 클래스 개수이다. 즉 S_k 각 클래스에 대한 영역분할 맵이다. 위의 수식을 통해 얻은 값을 기반으로 음향 이벤트 검출에 대한 성능을 측정하였다.

3.5 실험 결과

실험은 본 논문에서 제안하는 확장 합성곱 신경망 모델, 확장 합성곱 게이트 선형 유닛 모델과 베이스라인 모델인 영역분할 맵 추출 기반 음향 이벤트 검출 모델,[14] 그리고 기존에 제안되었던 합성곱 신경망을 이용하여 추출한 저차원 특징 맵을 고차원으로 확장한 FrameCNN[11]과 합성곱 신경망에서 주파수 축으로만 차원을 줄여 시간 축에 대한 정보는 유지하는 WLDCNN,[12] 합성곱 신경망과 완전연결 레이어에 게이트 선형 유닛을 적용한 Attention[13]을 가지고 음향 태깅과 음향 이벤트 검출에 대한 성능을 비교하였고, 잡음에 따른 모델의 성능 변화를 나타내었다.

기존의 베이스라인[16] 모델의 문제점으로 언급하였던 수용영역 부족에 대한 해결책으로 확장 합성곱 신경망을 적용하여 수용영역을 넓혀 준 확장 합성곱 신경망 모델의 성능을 보면 Tables 2 ~ 5에 나타내었듯이, 음향 태깅과 음향 이벤트 검출성능이 실험에 사용한 두 데이터베이스에서 베이스라인 모델보다 각각 6.38 %, 5.83 %와 7.94 %, 5.89 % 향상됨을 확인할 수 있다. 하지만 잡음 환경에 대한 성능 결과인 Tables 6과 7을 보면, 확장 합성곱 신경망 모델이 베이스라인 모델과 잡음이 가장 많은 환경인 SNR 0에서 F1 score가 비슷하거나 더 낮은 성능을 보이는 것을 확인하였다. 위의 현상은 Fig. 4에서 보이듯이 확장 모델이 수용영역이 넓어짐에 따라 더 넓은 범위의 이벤트에 대한 정보를 추출하지만, 넓은 범위의 잡음에 대한 정보 또한, 추출 되는 것으로 확인된다.

이에 앞에서 언급하였던 기존 모델의 문제점인 희소성에 대한 즉, 정보의 유효성 판단을 위해 확장 모델에 게이트 선형 유닛을 적용한 확장 게이트 선형 유닛 모델을 실험하였다. 확장 게이트 선형 유닛 모델은 Tables 2 ~ 5에 나타내었듯이, 조류 울음소리 데이터의 음향 태깅 성능을 제외한 모든 부분에서 가장 좋은 성능을 내는 것을 확인하였다. 또한, 정보에 대한 흐름을 제어함으로써 정보의 유효성을 판단할 수 있는 게이트 선형 유닛을 적용함으로 인해 잡음 환경 데이터에서도 확장 모델의 F1 score보다 음향 태깅에서 5.26 %, 음향 이벤트 검출에서 5.93 %의 성능 향상을 보였고, Fig. 4에서 확인할 수 있듯이 추출된 호랑지빠귀의 영역분할 맵에서 확장 모델보다 이벤트에 대한 정보가 다소 살아난 것을 볼 수 있고, 팔색조에 해당하는 영역분할 맵에서는 확장 모델에서 추출된 잡음에 대한 정보가 확장 게이트 선형 유닛 모델에서는 추출이 되지 않은 것을 확인할 수 있었다.

추가적으로 Table 3의 조류 울음소리의 음향 태깅 성능을 보면 Table 4의 음향 이벤트 검출 성능과는 다르게 확장 게이트 선형 유닛 모델 보다 확장 합성곱 신경망 모델의 성능이 더 높게 나온 것을 확인할 수 있다. 이 현상은 음향 태깅의 성능을 측정할 때 사용하는 예측값은 전체 픽셀의 정보를 이용하기 얻기 때문에, Fig. 4와 같이 확장 게이트 선형 유닛 모델을 통해 잡음에 대한 정보가 없어짐으로 음향 태깅의 성능이 다소 저하 될 수 있다. 즉, 영역분할 맵 추출 기반 음향 이벤트 검출에서는 음향 태깅의 성능이

Table 2. F1 score of audio tagging for URBAN-SED data.

Model	air conditioner	car horn	children-playing	dog bark	drilling	engine-idling	gun shot	jackhammer	siren	street music	Average
FrameCNN[13]	0.584	0.635	0.575	0.584	0.600	0.591	0.616	0.617	0.616	0.576	0.599
WLDCNN[14]	0.549	0.570	0.551	0.564	0.567	0.559	0.568	0.538	0.555	0.542	0.556
Attention[15]	0.549	0.570	0.551	0.564	0.567	0.559	0.568	0.538	0.555	0.542	0.556
Baseline[16]	0.601	0.637	0.613	0.601	0.660	0.614	0.622	0.627	0.613	0.604	0.619
Dilated	0.633	0.726	0.657	0.645	0.724	0.682	0.734	0.698	0.682	0.648	0.683
DCGLU	0.633	0.735	0.672	0.648	0.727	0.682	0.749	0.707	0.682	0.660	**0.690**

Table 3. F1 score of audio tagging for bird sound data.

Model	Terpsiphone atrocaudata	Zosterops japonicus	Lanius bucephalus	Falco peregrinus	Hypsipetes amaurotis	Pitta nympha	Zootheara dauma	Columba janthina	Average
FrameCNN[13]	0.484	0.275	0.278	0.413	0.734	0.743	0.788	0.3	0.502
WLDCNN[14]	0.484	0.283	0.278	0.413	0.734	0.743	0.788	0.3	0.503
Attention[15]	0.484	0.283	0.278	0.413	0.734	0.743	0.788	0.3	0.503
Baseline[16]	0.524	0.320	0.6	0.632	0.741	0.749	0.862	0.315	0.593
Dilated	0.604	0.320	0.9	0.806	0.768	0.829	0.903	0.25	**0.672**
DCGLU	0.745	0.409	0.671	0.8	0.763	0.798	0.898	0.255	0.667

Table 4. F1 score of sound event detection for URBAN-SED data.

Model	air condit-ioner	car horn	children-playing	dog bark	drilling	engine-idling	gun shot	jackham-mer	siren	street music	Average
FrameCNN[13]	0.200	0.147	0.204	0.183	0.211	0.212	0.153	0.223	0.228	0.202	0.196
WLDCNN[14]	0.156	0.100	0.169	0.156	0.120	0.154	0.127	0.108	0.157	0.158	0.140
Attention[15]	0.170	0.115	0.183	0.164	0.170	0.182	0.130	0.164	0.178	0.173	0.162
Baseline[16]	0.408	0.374	0.441	0.369	0.532	0.473	0.333	0.527	0.474	0.450	0.438
Dilated	0.428	0.462	0.490	0.433	0.552	0.542	0.450	0.582	0.531	0.495	0.496
DCGLU	0.429	0.464	0.505	0.441	0.559	0.558	0.469	0.594	0.544	0.518	**0.508**

Table 5. F1 score of sound event detection for bird sound data.

Model	Terpsiphone atrocaudata	Zosterops japonicus	Lanius bucephalus	Falco peregrinus	Hypsipetes amaurotis	Pitta nympha	Zootheara dauma	Columba janthina	Average
FrameCNN[13]	0.177	0.086	0.096	0.147	0.355	0.349	0.382	0.1	0.212
WLDCNN[14]	0.178	0.09	0.096	0.151	0.361	0.353	0.392	0.101	0.215
Attention[15]	0.179	0.09	0.097	0.151	0.360	0.353	0.392	0.101	0.215
Baseline[16]	0.487	0.319	0.600	0.717	0.501	0.589	0.789	0.188	0.524
Dilated	0.588	0.274	0.804	0.746	0.532	0.730	0.812	0.175	0.583
DCGLU	0.729	0.411	0.708	0.757	0.526	0.671	0.811	0.214	**0.603**

Table 6. Audio tagging for noise data.

	20 dB			10 dB			0 dB		
Model	F1	AUC	mAP	F1	AUC	mAP	F1	AUC	mAP
Baseline[16]	0.560	0.740	0.654	0.525	0.713	0.576	0.507	0.637	0.477
Dilated	0.586	0.797	0.737	0.562	0.766	0.663	0.482	0.675	0.542
DCGLU	**0.612**	**0.822**	**0.767**	**0.581**	**0.806**	**0.710**	**0.525**	**0.722**	**0.593**

Table 7. Sound event detection for noise data.

	20 dB			10 dB			0 dB		
Model	F1	AUC	mAP	F1	AUC	mAP	F1	AUC	mAP
Baseline[16]	0.369	0.727	0.425	0.326	0.718	0.404	0.271	0.671	0.280
Dilated	0.407	0.778	0.551	0.364	0.767	0.514	0.278	0.691	0.366
DCGLU	**0.453**	**0.803**	**0.576**	**0.410**	**0.789**	**0.555**	**0.338**	**0.724**	**0.430**

음향 이벤트 검출 성능에 영향을 미치기는 하지만, 음향 태깅 성능이 높다고 해서 음향 이벤트 검출 성능이 높은 것이 아니므로, 모델의 성능을 음향 태깅 성능을 기준으로 평가하기에는 적합하지 않은 것으로 보인다.

IV. 결 론

본 논문에서는 기존에 제안되었던 음향 이벤트 검출을 위한 시간-주파수 영역분할 맵 추출 기반 모델의 수용영역과 희소성 부족에 관한 문제를 적은 양의 파라미터로 효율적으로 수용영역을 넓혀 줄 수 있는 확장 합성곱 신경망과 정보의 흐름을 제어함으로써 정보의 유효성을 판단할 수 있는 선형 게이트 유닛를 적용한 확장 게이트 선형 유닛 모델을 제안함으로써 개선하였다. 또한, 이벤트가 겹치지 않은 데이터로 실험한 베이스라인과 달리 실제 상황에서와같이 이벤트가 겹쳐서 발생하는 데이터베이스인 URBAN-SED와 직접 제작한 조류 울음소리 데이터를 이용하여 실험하였고, 잡음 환경에 따른 실험도 진행하였다. 이에 제안하는 확장 게이트 선형 유닛 모델이 잡음 환경을 포함한 모든 데이터에서 가장 좋은 음향 이벤트 검출성능을 보이는 것을 확인하였다. 추후 본 모델은 강한 레이블에 비해 적은 정보를 가지고 있는 약한 레이블에서 최대한 많은 정보를 얻을 수 있도록 다양한 손실 함수를 구성하여 적용할 예정이며, 영역분할 맵 추출 방식은 글로벌 풀링에 의존되기 때문에, 새로운 글로벌 풀링 방법을 모색하여 모델을 개선할 예정이다.

감사의 글

본 결과물은 환경부의 재원으로 한국환경산업기술원의 환경정책기반공공기술개발사업의 지원을 받아 연구되었습니다(2017000210001).

References

1. P. K. Atrey, N. C. Maddage, and M. S. Kankanhlli, “Audio based event detection for multimedia surveillance,” Proc. IEEE ICASSP. 5. V-V (2006).
2. J. Maxime, X. Alameda-Pineda, L. Girin, and R. Horaud, “Sound representation and classification benchmark for domestic robots,” Proc. IEEE ICRA. 6285-6292 (2014).
3. D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, “Bird detection in audio: a survey and a challenge,” Proc. IEEE 26th MLSP. 1-6 (2016).
4. D. Stowell and M. D. Plumbley, “Audio-only bird classification using unsupervised feature learning,” Proc. CLEF. 673-684 (2014).
5. K. Ko, J. Park, D. K. Han, and H. Ko, “Channel and frequency attention module for diverse animal sound classification,” IEICE Trans. on Information and Systems, **E102-D**, 2615-2618 (2019).
6. S. Park, M. Elhilali, D. K. Han, and H. Ko, “Amphibian sounds generating network based on adversarial learning,” IEEE Signal Processing Letters, **27**, 640-644 (2020).
7. K. Ko, S. Park, and H. Ko, “Convolutional neural netework based amphibian sound classification using covariance and modulogram” (in Korean), J. Acoust. Soc. Kr. **37**, 61-65 (2018).
8. D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” IEEE Trans. Multimedia, **17**, 1733-1746 (2015).
9. G. Parascandolo, H. Huttunen, and T. Virtanen, “Recurrent neural networks for polyphonic sound event detection in real life recordings,” Proc. IEEE ICASSP. 6440-6444 (2016).
10. A. Mesaros, T. Heittola, and T. Virtanen, “TUT database for acoustic scene classification and sound event detection,” Proc. 24th EUSIPCO. 1128-1132 (2016).
11. S.-Y. Chou, J.-S. R. Jang, and Y.-H. Yang, “Frame CNN: A weakly-supervised learning framework for frame-wise acoustic event detetion and classification,” DACSE. Tech. Rep., 2017.
12. A. Kumar and B. Raj, “Deep cnn framework for audio event recognition using weak labeled web data,” arXiv: 1707.02530 (2017).
13. Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, “Large-scale weakly supervised audio classification using gated convolutional neural network,” Proc. IEEE ICASSP. 121-125 (2018).
14. Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, “Sound event detection and time-frequency segmentation from weak labelled data,” IEEE/ACM Trans. on Audio, Speech, And Lang. Processing, **27**, 777-787 (2019).
15. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,”

arXiv: 1409.1556 (2014).
16. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE CVPR. 770-778 (2016).
17. Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” Proc. PMLR. **70**, 933-941 (2017).
18. Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian, “Environmental sound classification with dilated convolutions,” Applied Acoustics, **148**, 123-132 (2019).
19. J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, “SCAPER:a library for soundscape synthesis and augmentation,” Proc. IEEE WASPAA. 344-348 (2017).
20. A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” Proc. ECCV. 695-711 (2016).
21. Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, “DCASE 2018 challenge baseline with convolutional neural networks,” DACSE. Tech. Rep., 2018.
22. K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Tuda, and K. Takeda, “Weakly-supervised sound event detection with self-attention,” Proc. IEEE ICASSP. 66-70 (2020).
23. Y. Li, M. Liu, K. Drossos, and T. Virtanen, “Sound event detection via dilated convolutional recurrent neural networks,” Proc. IEEE ICASSP. 286-290 (2020).
24. D. Kingma and J. Ba, “Adam: a method for stochastic optimization,” arXiv:1412.6980 (2015).
25. S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” Proc. 32nd ICML. 448-456 (2015).
26. J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” Radiology, **431**, 29-36 (1982).
27. R. Girshich, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” Proc. IEEE CVPR. 580-587 (2014).

저자 약력

▸ 박 충 호 (Chungho Park)

2019년 2월 : 백석대학교 멀티미디어공학 학사 취득
2019년 3월 ~ 현재 : 고려대학교 전기전자공학과 석사과정

▸ 김 동 현 (Donghyun Kim)

2018년 2월 : 영남대학교 전기공학과 학사 취득
2018년 3월 ~ 현재 : 고려대학교 전기전자공학과 석박사통합과정

▸ 고 한 석 (Hanseok Ko)

1982년 : Carnegie-Mellon Univ. 전기공학 공학사 취득
1988년 : Johns Hopkins Univ. 전자공학 공학석사 취득
1992년 : Catholic Univ. of America 전자공학 공학박사 취득
1994년 ~ 현재 : 고려대학교 전기전자공학과 교수