한국음향학회지 제40권 제5호 pp. 479~487 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.5 (2021)
https://doi.org/10.7776/ASK.2021.40.5.479

pISSN : 1225-4428
eISSN : 2287-3775

잡음 학생 모델 기반의 자가 학습을 활용한 음향 사건 검지

Sound event detection model using self-training based on noisy student model

김남균,[1] 박창수,[1] 김홍국,[1,2†] 허진욱,[3] 임정은[3]

(Nam Kyun Kim,[1] Chang-Soo Park,[1] Hong Kook Kim,[1,2†] Jin Ook Hur,[3] and Jeong Eun Lim[3])

[1]광주과학기술원 전기전자컴퓨터공학부, [2]광주과학기기술원 AI 대학원, [3]한화테크윈 AI연구소

(Received July 16, 2021; accepted August 17, 2021)

초 록: 본 논문에서는 잡음 학생 모델 기반의 자가 학습을 활용한 음향 사건 검지 기법을 제안한다. 제안된 음향 사건 검지 모델은 두 단계로 구성된다. 첫 번째 단계에서는 잔차 합성곱 순환 신경망(Residual Convolutional Recurrent Neural Network, RCRNN)을 훈련하여 레이블이 지정되지 않은 비표기 데이터셋의 레이블 예측에 활용한다. 두 번째 단계에서는 세 가지 잡음 종류를 적용한 잡음 학생 모델을 자가학습 기법으로 반복하여 학습한다. 여기서 잡음 학생 모델은 SpecAugment, Mixup, 시간-주파수 이동을 활용한 특징 잡음, 드롭아웃을 활용한 모델 잡음, 그리고 semi-supervised loss function을 적용한 레이블 잡음을 활용하여 학습된다. 제안된 음향 사건 검지 모델의 성능은 Detection and Classification of Acoustic Scenes and Events(DCASE) 2020 Challenge Task 4의 validation set으로 평가하였다. DCASE 2020 챌린지 데이터셋의 baseline 및 최상위 랭크된 모델과 이벤트 단위 F1 점수 성능을 비교한 결과, 제안된 음향 사건 검지 모델이 단일 모델과 앙상블 모델에서 최상위 모델 대비 F1 점수를 각각 4.6 %와 3.4 % 향상시켰다.
핵심용어: 음향 사건 검지, 자가 학습, 잡음 학생 모델, 준지도 손실함수

ABSTRACT: In this paper, we propose an Sound Event Detection (SED) model using self-training based on a noisy student model. The proposed SED model consists of two stages. In the first stage, a mean-teacher model based on an Residual Convolutional Recurrent Neural Network (RCRNN) is constructed to provide target labels regarding weakly labeled or unlabeled data. In the second stage, a self-training-based noisy student model is constructed by applying different noise types. That is, feature noises, such as time-frequency shift, mixup, SpecAugment, and dropout-based model noise are used here. In addition, a semi-supervised loss function is applied to train the noisy student model, which acts as label noise injection. The performance of the proposed SED model is evaluated on the validation set of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 Challenge Task 4. The experiments show that the single model and ensemble model of the proposed SED based on the noisy student model improve F1-score by 4.6 % and 3.4 % compared to the top-ranked model in DCASE 2020 challenge Task 4, respectively.
Keywords: Sound event detection, Self-training, Noisy student model, Semi-supervised loss function

PACS numbers: 43.60.Bf, 43.60.Lq

I. 서 론

소리는 일상생활에서의 중요한 정보를 포함하고 있으며 우리 주변에서 발생하는 개별 음향 사건에 따라 해당 장면을 이해하는 데 큰 도움을 준다.[1] 음향 장면 인지 분야는 기계학습 및 인공지능을 기반으로 다양한 알고리즘들이 연구되고 있으며, 다음향 환경에서의 음향 사건을 인지하는 음향 사건 검지

†**Corresponding author:** Hong Kook Kim (hongkook@gist.ac.kr)
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Republic of Korea
(Tel: 82-62-715-2228, Fax: 82-62-715-2204)

기술이 주목받고 있다. 특히 음향 사건 검지 기술은 음향 인지와 관련된 광범위 응용 분야에 활용될 수 있다. 특히 영상 사건 검지 기술은 날씨, 조도 및 사각지대 등과 같은 환경에서는 검지 불가능하다는 단점을 가지며, 음향 사건 검지 기술은 이러한 영상 사건 검지 기술과 연동되는 방향으로 활용될 수 있다.[1] 또한, 음향 사건 검지 기술은 유리 파손음, 총소리, 타이어 마찰음 또는 자동차 충돌음과 같은 물리적 사건을 검지할 수 있고,[2] 소셜 미디어 콘텐츠를 더 자세히 이해할 수 있는 오디오 자막생성,[3] 생활 지원 및 의료[4] 등에 활용될 수 있다.

음향 사건 검지 모델은 주로 입력 오디오 샘플에 대해 음향 이벤트 종류와 해당 이벤트의 시작 시점과 끝 시점 정보가 표기된 강력하게 레이블링된 데이터를 활용하여 학습된다. 지난 10년 동안 멜-주파수 켑스트럼 계수 기반의 음향 특징을 활용한 Support Vector Machine(SVM)[5] 및 Gaussian Mixed Model-Hidden Markov Model(GMM-HMM)[6,7]과 같이 기계학습 기반의 음향 사건 검지 모델이 제안되었다. 최근 음성인식, 음성합성 등에서 활용되는 심층 신경망 기반 모델을 응용한 음향 사건 검지 기술이 활발히 연구되고 있다. 완전 연결 신경망,[8] 합성곱 신경망(Convolutional Neural Network, CNN),[9,10] 순환 신경망(Recurrent Neural Network, RNN)[11] 및 합성곱 순환 신경망(Convolutional Recurrent Neural Network, CRNN)[12,13]과 같은 다양한 신경망 구조들이 음향 사건 검지 기술에 적용되었다.

앞서 설명한 음향 사건 검지 모델 학습은 강력하게 레이블링된 데이터를 많이 필요로 한다. 이러한 훈련 데이터는 실제 환경에서 수집된 데이터를 활용하여야 한다. 하지만, 이러한 데이터의 레이블을 생성하기 위해서는 큰 비용과 시간이 필요하다. 이에 대한 대안으로 오디오 샘플에 대해 시작 시점과 끝 시점 정보 없이 음향 이벤트 종류만 표기된 약하게 레이블링된 데이터를 결합하여 음향 사건 검지 모델 학습에 사용한다.[1]

강력하게 레이블링된 데이터와 약하게 레이블링된 데이터 외에도 레이블이 지정되지 않은 비표기 데이터(unlabeled data)를 사용하여 음향 사건 검지 성능 향상이 가능하다. 그 방법의 하나는 평균 교사 모델을 활용한 방법이다.[14] 평균 교사 모델 기반의 음향 사건 검지 모델은 학생 및 교사 두개의 모델이 있으며, 여기서 학생 모델은 교사 모델에 의해 예측되는 레이블과 일관성을 향상하는 방향으로 학습된다. 그런 다음, 각 epoch에 대한 학생 모델의 가중치 갱신에 따라 교사 모델도 업데이트된다. 특히 평균 교사 모델의 손실함수는 데이터 유형에 따라 구성한다. 즉 강력하게 레이블링된 데이터에 대한 손실함수는 강력한 레이블 예측값과 대상 레이블 간의 Binary Cross Entropy(BCE)로 학습된다. 또한, 약하게 레이블링된 데이터에 대한 손실함수는 약한 레이블 예측값과 대상 레이블 간의 BCE로 구성된다. 비표기 데이터 학습을 위한 손실함수는 학생 모델과 교사 모델의 예측 사이의 Mean Square Error(MSE)를 사용하여 학습한다. 이는 강력하게 레이블링된 데이터, 약하게 레이블링된 데이터, 그리고 비표기 데이터를 활용하여 학습된다. 이 평균 교사 학습 기반 모델은 Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 챌린지 Task 4의 baseline 모델로 제안되었으며, DCASE 2019 및 DCASE 2020 챌린지 Task 4에서 상위 순위를 달성한 모델에서 평균 교사 모델을 활용하였다.[15] 하지만, 평균 교사 모델의 예측은 완벽하지 않기 때문에, 약하게 레이블링된 데이터와 비표기 데이터에 대해 정확한 음향 종류와 시간 정보를 제공할 수 없다는 단점을 가진다.

이러한 문제를 해결하기 위해 본 논문에서는 강력하게 레이블링된 데이터, 약하게 레이블링된 데이터 및 비표기 데이터를 활용한 음향 사건 검지 기법을 제안한다. 합성곱 순환 신경망 구조의 합성곱 신경망을 잔차 학습을 통해 개선하였다. 또한, 같은 음향 사건에 속하는 샘플이라도 다른 특성을 가질 수 있으므로,[16] 이러한 특성을 학습하기 위하여 많은 합성곱 신경망 층을 활용하여야 한다. 이 경우, 각 합성곱 층은 잔차 학습을 사용하면 기울기 소실 문제를 극복하여 신경망을 더욱 효과적으로 학습할 수 있다. 결과적으로 합성곱 순환 신경망 대신 잔차 합성곱 순환 신경망(Residual Convolutional Recurrent Neural Network, RCRNN)을 활용한다.[17] 또한, 약하게 레이블링된 데이터와 비표기 데이터 활용을 위하여 잡음 학생 모델을 활용한 자가 학습 기반의 음향 사건 검지 기법을 제안한다.

먼저, RCRNN 기반의 평균 교사 모델을 활용하여 강력한 레이블링된 데이터, 약하게 레이블링된 데이터 그리고 비표기 데이터를 포함한 모든 데이터를 활용하여 학습한다. 다음으로, 잡음 학생 모델은 강하게 레이블링된 데이터는 대상 레이블을 사용하고, 약하게 레이블링된 데이터와 비표기 데이터에 대해 앞서 훈련된 RCRNN 기반의 평균 교사 모델을 활용하여 예측된 레이블을 활용하여 학습된다. 특히 self-training 기법[18]을 적용하여 잡음 학생 모델을 학습시킨다. 여기서 잡음 학생 모델은 특징 잡음, 모델 잡음, 그리고 레이블 잡음 세가지를 활용하였다. 특징 잡음의 경우, SpecAugment,[19] mixup,[20] 시간-주파수 이동[21]을 활용하였고, 드롭아웃기반의 모델 잡음, 레이블 잡음에 해당하는 semi-supervised loss function[17]을 활용하여 학습한다.

본 논문의 구성은 다음과 같다. 서론에 이어 II절에서 본 논문에서 사용되는 음향 사건 검지 분야의 데이터셋에 대해 기술한다. 이어서 III절에서는 잡음 학생 모델을 활용한 자가 학습 기반의 음향 사건 검지 기법을 제안한다. 그리고 IV절에서는 제안하는 음향사건 검지 기법의 성능평가 결과에 대해 기술한 후 V절에서는 본 논문의 결론을 맺는다.

II. DCASE 2020 챌린지 데이터셋

DCASE 2020 챌린지 데이터셋은 모델 학습을 위한 데이터셋으로 세가지 데이터셋을 제공한다.[14] 즉, 1) 약하게 레이블링된 데이터셋, 2) 레이블이 지정되지 않은 비표기 데이터셋, 그리고 3) 합성된 강력하게 레이블링된 데이터셋으로 구성된다. 1)과 2) 데이터셋의 경우, 실제 환경에서 수집된 데이터셋인 Audioset에서 가져온 데이터셋에 비해 강력하게 레이블링된 데이터셋은 Scaper soundscape 합성 및 증강 라이브러리를 활용하여 생성된다. 1), 2), 그리고 3) 데이터셋의 경우 각각 1,578개, 14,412개, 그리고 2,584개의 오디오 클립이 포함되어 있다. 또한, 모델 평가에 활용되는 데이터셋으로 공개 평가 데이터셋은 강력한 레이블을 포함한 1,168개의 validation 평가셋이 제공된다. 각각의 오디오 클립은 44.1 kHz로 샘플링되었으며, 최대 10 s 분량으로 구성된다.

주어진 데이터셋은 먼저 모노 채널로 다운 믹싱되고 44.1 kHz는 16 kHz로 다운샘플링된다. 다음으로, 입력 오디오 클립은 255개 샘플의 홉 길이가 있는 2048개 샘플의 연속 프레임으로 분할된다. 그리고 나서, 분리된 각 신호에 2048-point 고속푸리에 변환에 대해 128차원 mel-filterbank 분석을 수행한다. 각 10 s 오디오 클립은 628개의 프레임으로 표현되어 음향 사건 검지 모델의 입력 특성의 크기는 1 × 628 × 128이 된다. 이때, 10 s보다 짧은 오디오 클립에는 zero padding이 적용된다. 마지막으로 추출된 멜 스펙트로그램은 모든 학습 오디오 클립에 대한 전역 평균과 표준 편차로 정규화된다.

III. 제안된 음향 사건 검지 모델

3.1 RCRNN 기반 평균 교사 모델 기반 음향 사건 검지

Fig. 1과 같이 제안된 음향 사건 검지 모델의 첫 번째 단계는 RCRNN 기반 평균 교사 모델을 기반으로 구성되며,[17] 이는 CRNN 기반의 평균 교사 모델[14]을 RCRNN으로 대체한 동일한 모델 구조를 가진다. Table 1은 평균 교사 모델에 사용된 RCRNN의 구조와 파라미터를 보여 준다.

먼저 628개 프레임의 특징을 그룹화하여 (628 × 128) 차원의 스펙트럼 이미지를 만든 다음, RCRNN의 입력 특징으로 사용한다. Table 1에서 설명한 바와 같이 RCRNN의 합성곱 블록은 1개의 stem block과 5개의 잔차 컨볼루션 블록으로 구성되며, 여기서 stem block은 첫 번째 및 두 번째 컨볼루션 블록에 대해 각각 16개 및 32개의 커널을 갖는 2개의 컨볼루션 블록으로 구성된다. 각 컨볼루션 블록에는 stride가 (1 × 1)인 (7 × 7) 합성곱 커널이 있으며 배치 정규화(Batch Normalization, BN), 게이트 선형 유닛(Gated Linear Unit, GLU) 활성화 및 (2 × 2) 평균 풀링 레이어가 연결된다. 이는 입력 시간 축을 1/2로 다운샘플링하는 역할을 수행한다.

다음으로, 각 잔차 컨볼루션 블록의 출력에 합성곱 블록 주의 모듈(Convolutional Block Attention Module, CBAM)[22]를 적용한다. 합성곱 레이어를 모두 마친 후 (128 × 157 × 1) 차원의 특징맵이 순환 블록에 적용

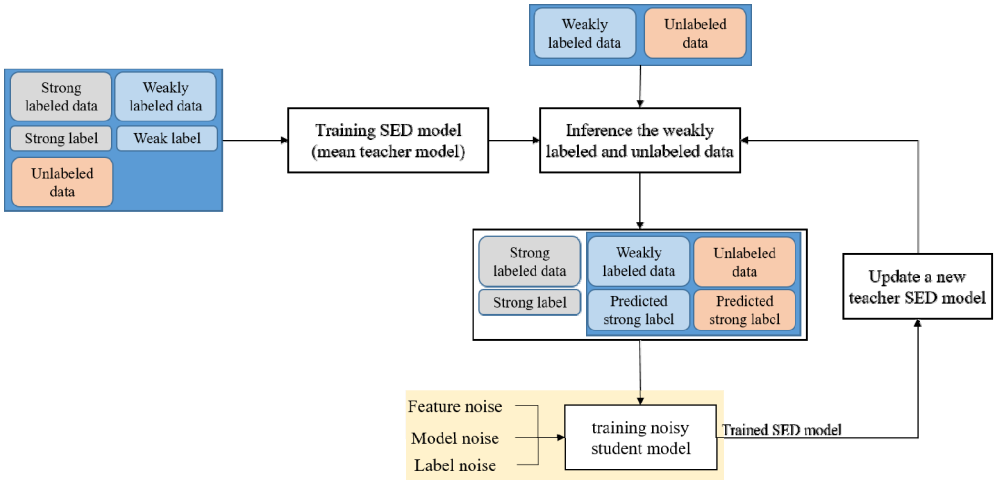

Fig. 1. (Color available online) Training procedure of the proposed sound event detection model composed of the RCRNN-based mean-teacher model for predicting strong labels and the self-trained noisy student model with noise injections and a semi-supervised loss function.

Table 1. Network architecture of a residual convolutional neural network in the RCRNN used in the mean-teacher model.

Name	Layers	Output shape
Input layer	Input: log-mel spectrogram	1×628×128
Stem block	$\begin{pmatrix}7\times7,\ Conv2D,\ @16,\\ GLU,\ BN\end{pmatrix}$ 2×2 average pooling layer	16×314×64
	$\begin{pmatrix}7\times7,\ Conv2D,\ @32,\\ GLU,\ BN\end{pmatrix}$ 2×2 average pooling layer	32×157×32
Residual convolution block	$\begin{pmatrix}3\times3,\ Conv2D,\ @64,\\ ReLU,\ BN\end{pmatrix}\times2$ Self-attention module (CBAM) 1×2 average pooling layer	64×157×16
	$\begin{pmatrix}3\times3,\ Conv2D,\ @128,\\ ReLU,\ BN\end{pmatrix}\times2$ Self-attention module (CBAM) 1×2 average pooling layer	128×157×8
	$\begin{pmatrix}3\times3,\ Conv2D,\ @128,\\ ReLU,\ BN\end{pmatrix}\times2$ Self-attention module (CBAM) 1×2 average pooling layer	128×157×4
	$\begin{pmatrix}3\times3,\ Conv2D,\ @128,\\ ReLU,\ BN\end{pmatrix}\times2$ Self-attention module (CBAM) 1×2 average pooling layer	128×157×2
	$\begin{pmatrix}3\times3,\ Conv2D,\ @128,\\ ReLU,\ BN\end{pmatrix}\times2$ Self-attention module (CBAM) 1×2 average pooling layer	128×157×1
Recurrent block	128 BiGRU cells ×2	256×157

된다. 순환 블록은 입력 특징의 시간 정보를 학습하기 위한 두 개의 양방향 게이트 순환 유닛(Bidirectional Gated Recurrent Unit, BiGRU)로 구성되며 각 GRU에 대한 활성화 함수로 정류 선형 유닛(Rectified Linear Unit, ReLU)이 사용된다. 순환 블록의 (256 × 157) 차원의 출력은 fully connected layer에 연결되고 시그모이드 함수를 적용한 후 (157 × 10) 차원으로 출력된다. 여기서 10은 감지할 음향 사건의 수를 나타낸다. (157 × 10) 차원 출력은 음향 사건 종류 및 시간 정보를 포함하는 강력한 레이블과 연관된다. 또한, 가중치 풀링 레이어가 (157 × 10) 차원 출력에 적용되어 주어진 오디오 클립에 대한 약한 레이블을 예측하는 (1 × 10) 차원 출력을 얻는다.

지금까지 훈련된 RCRNN 기반 평균 교사 모델은 약하게 레이블링된 데이터셋과 레이블이 없는 데이터셋의 강력한 레이블 예측값을 생성하는 데 사용된다. 여기서, 예측된 강력한 레이블은 임계값이 0.5로 시그모이드 출력에 임계값을 적용하여 생성한다. 이러한 예측된 레이블을 사용하여 다음 세부 절에서 설명할 잡음 학생 모델 기반 음향 사건 검지 모델을 학습에 사용된다.

3.2 잡음 학생 모델 기반 음향 사건 검지 모델

제안된 음향 사건 점지 모델의 두 번째 단계는 RCRNN 기반의 잡음 학생 모델이다. 잡음 학생 모델

을 훈련하기 위하여 앞서 설명한 RCRNN 기반 평균 교사 모델에서 예측한 강력한 레이블은 약하게 레이블링된 데이터셋과 비표기 데이터셋에 활용되고, 강력하게 레이블링된 데이터셋은 대상 레이블을 그대로 사용된다.

잡음 학생 모델을 훈련하기 위해 제안된 음향 사건 검지 모델의 첫 번째 단계에서 평균 교사 모델에서 예측된 강한 레이블은 약하게 레이블링된 데이터셋 또는 비표기 데이터셋에 사용되는 반면 강한 레이블이 지정된 데이터에는 지정된 강한 레이블이 사용된다. 그 후 SpecAugment,[19] mixup,[20] 시간-주파수 이동[21]의 시간-주파수 마스킹의 특징 잡음 기술을 순차적으로 적용하여 입력 스펙트럼 이미지에 적용된다. 여기서, 시간-주파수 마스킹은 시간 및 주파수 영역의 값을 0으로 대체하여 작동하고, mixup 기법은 입력 특징을 현재 입력 특징과 다른 특징을 혼합하여 잡음 데이터를 생성한다. 시간 주파수 이동은 주파수 및 시간 축에 대해 각각 평균이 0이고 표준 편차가 각각 4와 32인 임의의 가우스 잡음에 대해 시간 및 주파수 축을 따라 입력 스펙트럼 이미지를 순환 이동으로 적용한다. 또한, 잡음 학생 모델에 대한 모델 잡음 구현을 위해 0.5 확률의 드롭아웃을 적용한다. 마지막으로, 레이블 잡음 적용으로 semi-supervised loss function[17]이 사용된다.

본 논문에서의 semi-supervised loss function은 다음 식과 같이 정의된다.

$$L_{semi} = \sum_{i \in S} BCE(i;\theta) + \sum_{i \in W,U} BCE_{soft}(i;\theta), \qquad (1)$$

여기서 S, W 및 U는 강하게 레이블링된 데이터셋, 약하게 레이블링된 데이터셋 및 비표기 데이터셋을 각각 나타낸다. Eq. (1)에서 θ는 RCRNN 기반의 잡음 학생 모델을 나타낸다. 또한 $BCE(i;\theta)$는 BCE 손실함수이고 $BCE_{soft}(i;\theta)$는 RCRNN 기반 평균 교사 모델의 이진화된 강력한 레이블과 θ의 예측 출력 사이의 BCE 손실함수로 정의된다. 즉, $BCE_{soft}(i;\theta)$는 다음 식과 같이 정의된다.

$$BCE_{soft}(i;\theta) = -(\bar{y}_i \log \hat{y}_{i,\theta} + (1-\bar{y}_i)\log(1-\hat{y}_{i,\theta})), \qquad (2)$$

여기서 $\hat{y}_{i,\theta}$는 i번째 오디오 클립에 대한 RCRNN 기반 잡음 학생 모델 θ의 예측값이다. Eq. (2)에서 $\bar{y}_i$는 이진화된 강력한 레이블 사이의 보간될 대상이며 다음 식과 같이 계산된다.

$$\bar{y}_i = \beta \hat{y}_{i,\theta} + (1-\beta)\hat{y}_{i,\theta_m}, \qquad (3)$$

여기서 $\hat{y}_{i,\theta_m}$는 RCRNN 기반 평균 교사 모델 θ_m의 강력한 레이블 예측값이다. Eq. (3)에서 β는 손실함수에 대한 하이퍼파라미터로, 서로 다른 앙상블 모델을 얻기 위해 설정된다. 결과적으로 잡음이 있는 입력 스펙트럼 이미지를 사용하여 잡음이 있는 학생 모델은 특징잡음, 드롭아웃 및 semi-supervised loss function으로 학습된다.

잡음 학생 모델 훈련을 마친 후 모델 매개 변수는 평균 교사 모델의 교사 모델이 된다. 그런 다음 약하게 레이블이 지정되거나 비표기 데이터에 대한 강력한 레이블이 교사모델에 의해 업데이트되며, 이는 또한 잡음 학생 모델에 새 대상 레이블로 사용된다. 결론적으로, 잡음 학생 모델을 훈련하고 레이블을 업데이트하는 이 절차를 두 번 더 반복한다.

IV. 성능 평가

4.1 실험 환경

평균 교사 모델의 학습은 DCASE 2020 챌린지 Task 4의 baseline[14]의 학습 방식을 따라 RCRNN 모델을 학습하였다. 즉, 신경망 가중치는 Xavier 초기화를 사용하여 초기화되었으며 bias 값은 모두 0으로 초기화되었다. 다음으로, 드롭아웃은 0.5의 비율로 적용되었고, ADAM 최적화 알고리즘을 활용하여 모델을 학습하였다. 또한 50 epoch 후에 최대 학습률이 0.001에 도달하는 ramp-up 방식에 따라 학습률을 설정했다. 데이터 증대를 위해 시간-주파수 이동[21]과 mixup[20]을 사용하였다. 여기서 훈련에 사용된 배치의 수는 32로 설정하여 훈련하였다.

다음으로, 잡음 학생 모델은 훈련 셋의 모든 데이터를 5-fold로 나누고 5-fold 중 4개의 fold를 학습에 사용하는 5-fold cross validation을 기반으로 3.2 절에 설

명된 바와 같이 학습되었다. 다른 1개의 fold는 모델 검증에 사용되었다. 여기서, 학습률은 초기에 0.001로 설정되었으며 교차 검증에 활용된 손실함수 값을 판단하여 학습률을 감소시켰다.

4.2 평가 지표

제안된 음향 사건 검지 모델의 성능은 F1-score, Error Rate(ER)와 같은 객관적인 척도로 측정되었다. F1 점수는 다음 식과 같이 정의된다.[23]

$$\mathrm{Precision} = \frac{TP}{TP+FP}. \tag{4}$$

$$Recall = \frac{TP}{TP+FN}. \tag{5}$$

$$F1 = \frac{2 \bullet \mathrm{Precision} \bullet Recall}{\mathrm{Precison} + Recall}, \tag{6}$$

여기서 True Positive(TP), False Positive(FP), 그리고 False Negative(FN)은 각각 검지된 정답, 검지된 오답, 그리고 검지되지 않은 정답의 수이다. F1 점수가 높을수록 음향 사건 검지에 대한 더 나은 검지 성능을 의미한다. 이벤트 단위 F1-score는 각 이벤트별로 시작과 끝 지점을 일정 범위(본 실험에서는 0.2 s으로 정의) 내로 구간을 맞춘 수를 위 F1-score에 따라 계산된다. 본 계산 방식은 DCASE 2020 챌린지 베이스라인의 측정방법과 동일하게 측정하였다.[14]

ER은 삽입(I), 삭제(D), 치환(S)의 오류 수를 측정하여 다음 식과 같이 정의하였다.[23]

$$ER = \frac{\sum_i S(i) + \sum_i D(i) + \sum_i I(i)}{\sum_i N(i)}, \tag{7}$$

여기서 $S(i)$, $I(i)$, $D(i)$, $N(i)$은 각각 i번째 오디오 클립에서 삽입, 삭제, 대체 및 실측 음향 사건의 수를 의미한다. 따라서 낮은 ER은 더 나은 음향 사건 검지 성능을 나타낸다.

4.3 결과 비교

본 세부 절에서는 제안된 잡음 학생 모델 기반의 음향 사건 검지 모델을 DCASE 2020 챌린지 Task 4에 적용하여 챌린지 baseline 및 최상위 모델과 성능을 비교하였다. Table 2는 단일 모델 기준으로 DCASE 2020 Challenge Task 4 baseline, DCASE 2020 Challenge Task 4의 최상위 모델, RCRNN 기반 평균 교사 모델 및 제안된 잡음 학생 모델 기반의 음향 사건 검지 모델의 이벤트 기반 F1-score와 ER을 비교를 보여 준다. 표에서 보는 바와 같이, RCRNN 기반 평균 교사 모델은 DCASE 챌린지의 baseline 및 상위 모델보다 더 높

Table 2. Comparison of F1-score and ERs between the top-ranked sound event detection model and the proposed RCRNN-based noisy student sound event detection model.

Model	Event-based F1-score	ER
Baseline of DCASE 2020 Task 4[14]	34.8	-
Top-ranked model of DCASE 2020 Task 4[24]	46.0	-
RCRNN-based mean-teacher model	46.8	1.13
RCRNN, noisy student, $\beta = 1.0$ (without label noise)	50.9	1.00
RCRNN, noisy student, $\beta = 0.3$	50.8	0.97
RCRNN, noisy student, $\beta = 0.5$	51.2	0.96
RCRNN, noisy student, $\beta = 0.7$	51.4	0.96
RCRNN, noisy student, $\beta = 0.9$	50.1	0.98

Table 3. Comparison of F1-score and ERs between the top-ranked ensemble sound event detection model and the proposed RCRNN-based noisy student ensemble sound event detection model.

Model	Event-based F1-score	ER
Top-ranked model of DCASE 2020 Task 4 (6 model ensemble)[24]	50.6	-
RCRNN, noisy student, $\beta = 1.0$ (5 model ensemble)	52.6	0.92
RCRNN, noisy student, $\beta = 0.3$ (5 model ensemble)	51.7	0.94
RCRNN, noisy student, $\beta = 0.5$ (5 model ensemble)	52.7	0.93
RCRNN, noisy student, $\beta = 0.7$ (5 model ensemble)	54.0	0.89
RCRNN, noisy student, $\beta = 0.9$ (5 model ensemble)	51.8	0.93

Table 4. Ablation study for the proposed noisy student sound event detection model using an RCRNN-based teacher model with different types of noise injections.

Model	Feature noise	Model noise	Label noise	Event-based F1-score	Error rate
Baseline: CRNN-based mean-teacher model[16] (single model)	-	-	-	34.8	
RCRNN-based mean-teacher model (single model)	-	-	-	46.8	1.13
Noisy student sound event detection model (RCRNN model, single model)	-	√	√	46.8	1.05
	√	-	√	49.8	0.99
	√	√	-	50.9	1.00
	√	√	√	51.4	0.96

은 F1 점수를 달성하였다. 특히, 잡음 학생 모델 기반의 음향 사건 검지 모델은 β값과 관계없이 평균 교사 모델 대비 향상된 F1 점수를 얻었으며, 특히 $\beta = 0.7$ 일 때 4.6 % 개선으로 최고의 성능을 보였다.

Table 3은 제안된 잡음 학생 기반 음향 사건 검지 모델의 앙상블 버전과 DCASE 2020 챌린지의 최상위 버전 모델을 비교하였다. 본 논문에서의 모델 앙상블로서 각 5-fold cross validation 시 학습된 모델의 결과를 앙상블하였다. 즉, 5개의 모델을 앙상블하여 단일 모델 대비 성능을 개선하였다. 표에서 보는 바와 같이, 제안된 모델의 앙상블 버전과 DCASE 2020의 최상위 모델을 비교하였을 때, 상위 모델 대비 F1 점수를 3.4 % 증가시켰다.

마지막으로, Table 4는 제안된 잡음 학생 모델의 ablation study를 수행한 결과이다. 본 실험은 잡음 학생 모델을 구성하는 세가지 잡음을 한가지씩 제외하며 실험하여 어떤 잡음이 성능에 큰 영향을 미치는지 분석하였다. 분석한 결과, 특징잡음 유무에 따라 성능이 크게 개선됨을 확인하였으며, 모든 잡음을 추가하였을 때 음향 사건 검지 성능이 제일 개선됨을 확인할 수 있었다.

V. 결 론

본 논문에서는 잡음 학생 모델 기반의 음향 사건 검지 모델을 제안하였다. 제안된 음향 사건 검지 모델은 약하게 레이블링된 데이터와 비표기 데이터와 같은 훈련 데이터셋을 활용한 자가 학습을 기반으로 하였다. 특히, RCRNN 기반 평균 교사 모델을 사용하여 약하게 레이블링된 데이터셋과 비표기 데이터셋에서 각 오디오 클립의 대상 레이블을 예측하였다. 잡음 학생 모델 기반의 자가 학습을 구현을 위해 잡음 학생 모델은 데이터 증대 기반의 특징 잡음, 드롭아웃 기반 모델 잡음, semi-supervised loss function 기반의 레이블 잡음을 모델에 주입하였다. 특히, semi-supervised loss function의 cross-validation과 다른 하이퍼파라미터 값에 따라 잡음 학생 모델을 학습시켰고, 다섯가지의 서로 다른 모델을 앙상블 모델로 하여 성능을 개선하였다. 제안된 음향 사건 검지 모델의 성능은 DCASE 2020 챌린지 Task 4의 validation set에서 평가되었다. 결과적으로, DCASE 챌린지 Task 4의 baseline 모델과 최상위 모델과 성능을 비교하였을 때 높은 성능을 달성하였다. 특히 최상위 모델 대비 단일모델에서 4.6 %, 앙상블 모델에서 3.4 % F1 점수 향상을 보였다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00014, 재난상황 대응을 위한 엣지컴퓨팅 기반 시청각 인지지능 솔루션 개발).

References

1. T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events* (Springer, Heidelberg, 2018), Chap. 1.
2. J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing and mitigating urban noise pollution," Commun. ACM. **62**, 68-77 (2019).
3. K. Drossos, S. Adavanne, and T. Virtanen, "Automated

audio captioning with recurrent neural networks," Proc. IEEE WASPAA. 374-378 (2017).
4. Y. Zigel, D. Litvak, and I. Gannot, "A method for automatic fall detection of elderly people using floor vibrations and sound -Proof of concept on human mimicking doll falls," IEEE Trans. Biomed. Eng. **56**, 2858-2867 (2009).
5. A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," Pattern Recognit. Lett. **30**, 1281-1288 (2009).
6. A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," Proc. EUSIPCO. 1267-1271 (2010).
7. T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," EURASIP J. Audio, Speech, and Music Process. **2013**, 1-13 (2013).
8. E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," Proc. IJCNN. 1-7 (2015).
9. H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," Proc. IEEE ICASSP. 559-563 (2015).
10. H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," Proc. Interspeech, 3653-3657 (2016).
11. 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).
12. E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," IEEE/ACM Trans. on Audio, Speech, Lang. Process. **25**, 1291-1303 (2017).
13. S. Adavanne, P. Pertila, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," Proc. IEEE ICASSP. 771-775 (2017).
14. N. Turpault, R. Serizel, A. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," Proc. Workshop on DCASE. 253-257 (2019).
15. N. Turpault, R. Serizel, S. Wisdom, H. Erdogan, J. R. Hershey, E. Fonseca, P. Seetharaman, and J. Salamon, "Sound event detection and separation: A benchmark on DESED synthetic soundscapes," Proc. IEEE ICASSP. 840-844 (2021).
16. 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).
17. N. K. Kim and H. K. Kim, "Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function," IEEE Access, **9**, 7564-7575 (2021).
18. Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," Proc. IEEE/CVF CVPR. 10687-10698 (2020).
19. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint, arXiv:1904.08779 (2019).
20. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," arXiv preprint, arXiv:1710.09412 (2017).
21. L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4," DCASE 2019 Challenge, Tech. Rep., 2019.
22. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," Proc. ECCV. 3-19 (2018).
23. A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," Appl. Sci. **6**, 162-178 (2016).
24. K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda "Convolution augmented transformer for semi-supervised sound event detection," Proc. Workshop on DCASE. 100-104 (2020).

저자 약력

▸ 김 남 균 (Nam Kyun Kim)

2014년 2월 : 전남대학교 전자공학과 학사
2015년 8월 : 광주과학기술원 정보통신공학부 석사
2021년 8월 : 광주과학기술원 전기전자컴퓨터공학부 박사
2021년 7월 ~ 현재 : 한국자동차연구원 선임연구원

▸ 박 창 수 (Chang-Soo Park)

1979년 2월 : 한양대학교 전자공학과 학사
1981년 2월 : 서울대학교 전자공학과 석사
1990년 12월: Texas A&M University 전기전자공학과 박사
1982년~1999년 : ETRI 책임연구원
2007년~2008년 : University of California at Irvine, Visiting Professor
2000년 1월 ~ 2020년 2월 : 광주과학기술원 전기전자컴퓨터공학부 교수
2020년 3월 ~ 현재 : 광주과학기술원 전기전자컴퓨터공학부 명예교수

▸ 김 홍 국 (Hong Kook Kim)

1988년 2월 : 서울대학교 제어계측공학과 학사
1990년 2월 : 한국과학기술원 전기 및 전자공학과 석사
1994년 8월 : 한국과학기술원 전기 및 전자공학과 박사
1990년 ~ 1998년 : 삼성종합기술원 전문연구원
1998년 ~ 1998년 : MMC Technology 선임연구원
1998년 ~ 2003년 : AT&T Labs- Research Senior Member Technical Staff
2014년 ~ 2015년 : City University of New York, Visiting Professor
2003년 8월 ~ 현재 : 광주과학기술원 전기전자컴퓨터공학부/AI대학원 교수

▸ 허 진 욱 (Jin Ook Hur)

1999년 2월 : 고려대 산업공학과 석사
2001년 2월 : 고려대 산업공학과 박사수료
2001년 3월 ~ 2005년 6월 : ㈜아이디스 선임연구원
2006년 9월 ~ 2009년 12월 : 삼성전자 책임연구원
2010년 1월 ~ 2011년 9월 : 삼성테크윈 책임연구원
2011년 10월 ~ 현재 : 한화테크윈 AI연구소 수석연구원

▸ 임 정 은 (Jeong Eun Lim)

2001년 2월 : 연세대 전기및전자공학과 석사
2004년 8월 : 연세대 전기및전자공학과 박사
2004년 9월 ~ 2009년 12월 : 삼성전자 책임연구원
2010년 1월 ~ 2011년 9월 : 삼성테크윈 책임연구원
2011년 10월 ~ 현재 : 한화테크윈 AI연구소 수석연구원