한국음향학회지 제40권 제2호 pp. 139~147 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.2 (2021)
https://doi.org/10.7776/ASK.2021.40.2.139

pISSN : 1225-4428
eISSN : 2287-3775

평균-교사 합성곱 순환 신경망 모델을 이용한 약지도 음향 이벤트 검출 시스템의 성능 분석

Performance analysis of weakly-supervised sound event detection system based on the mean-teacher convolutional recurrent neural network model

이석진[1†]

(Seokjin Lee[1†])

[1]경북대학교 전자공학부, 경북대학교 전자전기공학부

(Received January 18, 2021; accepted February 4, 2021)

초 록: 본 논문은 데이터의 일부만 레이블링이 되어있는 약지도 학습을 기반으로 하는 음향 이벤트 검출 시스템을 소개 및 구현하고, 시뮬레이션을 통해 각 파라미터가 성능에 미치는 영향을 분석하였다. 음향 이벤트 검출 시스템은 음향 신호 내에 존재하는 이벤트의 종류, 시작/종료 시점을 추정하는 시스템으로, 이를 학습시키기 위해서는 음향 이벤트 신호와 그 종류, 시작/종료 시점에 대한 모든 정보가 제공되어야 한다. 하지만 이를 모두 표기하여 학습데이터를 만드는 것은 매우 큰 비용이 들어가며, 특히 시작/종료 시점을 정확히 표기하는 것은 매우 어렵다. 따라서 본 논문에서 다루는 약지도 학습 문제에서는 이벤트의 종류와 시작/종료 시점이 모두 표기된 "강하게 표기된 데이터"와, 이벤트의 종류만 표기된 "약하게 표기된 데이터", 그리고 아무런 표기가 되어 있지 않은 "미표기 데이터"를 이용하여 음향 이벤트 검출 시스템을 학습시킨다. 최근 이러한 문제에서는 평균-교사 모델을 이용한 음향 이벤트 검출 시스템의 성능이 우수하며, 따라서 널리 사용되고 있다. 다만, 평균-교사 모델은 많은 파라미터를 가지고 있고, 이는 성능에 영향을 다소 미칠 수 있으므로 신중하게 선택되어야 한다. 본 논문에서는 DCASE 2020 Task 4의 데이터를 이용하여 특징 값의 종류, 이동 평균 파라미터, 일관성 비용함수의 가중치, 램프-업 길이, 그리고 최대 학습율 등 5가지의 값에 대해 성능 분석을 진행하였으며, 각 파라미터에 대한 영향 및 최적 값에 대해 고찰하였다.

핵심용어: 음향 이벤트 검출, 약지도 학습, 평균-교사 모델, 합성곱 순환 신경망

ABSTRACT: This paper introduces and implements a Sound Event Detection (SED) system based on weakly-supervised learning where only part of the data is labeled, and analyzes the effect of parameters. The SED system estimates the classes and onset/offset times of events in the acoustic signal. In order to train the model, all information on the event class and onset/offset times must be provided. Unfortunately, the onset/offset times are hard to be labeled exactly. Therefore, in the weakly-supervised task, the SED model is trained by "strongly labeled data" including the event class and activations, "weakly labeled data" including the event class, and "unlabeled data" without any label. Recently, the SED systems using the mean-teacher model are widely used for the task with several parameters. These parameters should be chosen carefully because they may affect the performance. In this paper, performance analysis was performed on parameters, such as the feature, moving average parameter, weight of the consistency cost function, ramp-up length, and maximum learning rate, using the data of DCASE 2020 Task 4. Effects and the optimal values of the parameters were discussed.

Keywords: Sound event detection, Semi-supervised learning, Mean-teacher, Convolutional recurrent neural network

PACS numbers: 43.60.Bf, 43.60.Lq

†**Corresponding author:** Seokjin Lee (sjlee6@knu.ac.kr)
School of Electronics Engineering, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Republic of Korea
(Tel: 82-53-950-5523, Fax: 82-53-950-5505)

I. 서 론

최근 딥러닝 등 기계 학습 기법이 크게 발전함에 따라, 기계가 음향 신호를 인식 및 분석하는 기법에 대한 연구도 크게 발전하고 있다. 여러 연구자들은 단순히 성능이 좋은 알고리즘들을 개발하는 것에서 나아가, 실생활에 적용될 수 있는 다양한 문제들을 발굴하고 이를 풀기위한 노력들을 계속하고 있다.

현재 연구되고 있는 음향 인식 관련 연구들은 매우 다양하지만, 대부분의 과업들은 음향 환경 분류(Acoustic Scene Classification, ASC)[1] 및 음향 이벤트 검출(Sound Event Detection, SED)[2] 문제에서 파생되었다. 음향 환경 분류 문제와 음향 이벤트 검출 문제 또한 엄밀하게 구분되지는 않으나, 음향 환경 분류는 주로 10초 정도의 긴 음향 신호를 분류하는 문제를 일컫는 반면, 음향 이벤트 검출 문제는 짧은 음향 신호의 구간을 검출하고 이를 분류하는 문제를 말한다.[3]

많은 기계학습 기반 시스템과 마찬가지로, 음향 이벤트 검출 시스템을 학습시키기 위해서는 매우 많은 수의 데이터와 이에 대한 레이블 표기 작업이 필요하다. 기계 학습 기반의 음향 인지 시스템을 실제 응용 시스템에 적용시킬 때 매우 큰 걸림돌이 되는 부분이다. 특히, 음향 이벤트 검출 시스템은 음향 신호의 종류를 분류하고, 이벤트의 시작/종료 시점을 찾는 것을 목표로 하므로, 학습을 위해 제공되는 데이터 또한 음향 이벤트의 종류뿐만 아니라 시작/종료 시점에 대한 데이터를 가지고 있어야 한다. 당연히 이를 위해서는 더 많은 인력을 필요로 하는 문제가 발생한다. 이를 해결하기 위하여, 최근 음향 이벤트 검출 과업을 다루는 커뮤니티에서는 적은 수의 레이블 표기 데이터와 많은 수의 미표기 데이터를 이용하는 약지도 문제를 다루고 있다.[4]

음향 환경 분류 및 음향 이벤트 검출 문제를 다루는 커뮤니티 중 하나인 Detection and Classification of Acoustic Scenes and Events(DCASE) 에서는 음향 이벤트 검출을 위한 데이터베이스를 다음과 같이 세 개의 그룹으로 나누고 있다.[4] 하나는 강하게 표기된 데이터(strongly-labeled data)로, 이 종류의 학습 데이터는 각 클립이 포함하고 있는 음향 이벤트들의 종류 및 시작/종료 시점의 레이블을 모두 포함하고 있다. 다음은 약하게 표기된 데이터(weakly-labeled data)로, 이 종류의 학습 데이터는 포함된 음향 이벤트들의 종류만이 표기되어 있고, 존재 구간에 대한 정보는 포함되어 있지 않다. 마지막은 미표기된 데이터(unlabeled data)로, 이 종류의 학습 데이터는 어떠한 레이블링 없이 오디오 클립만으로 이루어져 있다.

이와 같이 레이블링 없는 미표기 데이터를 포함하는 학습 기법을 약지도 학습이라 하며, 이를 위해서 다양한 학습 기법들이 연구된 바 있다. 이러한 기법들은 주로 잡음이 있는 교사 모델을 이용하였을 때 더욱 향상된 결과가 나온다는 관찰에 근거하여 개발되었다. 이는 의사-앙상블[5] 기법에서 도입되었으며, 이후 영상 분류 문제에 적용되어 π모델이라 명명되었다.[6] 이와 같이 개발된 π모델은 각 학습 샘플의 예측 값을 지수 가중치를 이용하여 평균을 취하는 시간 앙상블[7] 기법으로 발전되었다. 최근에는, 이를 교사-학생 모델을 이용하여 변형한 평균-교사 모델이 고안되었다.[8]

영상 분류 문제에서 먼저 적용되었던 평균-교사 모델은 이후 변형되어 약지도 음향 이벤트 검출 문제에 성공적으로 적용되었으며,[9] 최근 DCASE 등의 대회에서 좋은 성능을 보이고 있는 약지도 음향 이벤트 검출 기법들의 핵심 기법으로 사용되고 있다.[10,11] 이와 같이 평균-교사 모델이 약지도 음향 이벤트 검출에 성공적으로 사용되고 있는 반면, 기존의 모델보다 다소 복잡한 만큼 여러 파라미터를 가지고 있고, 이를 주의깊게 설정해 주어야 한다. 데이터의 레이블링이 완벽하지 않은 약지도 학습의 특성상, 검출 성능이 이러한 설정 값에 다소 민감하게 반응하게 된다. 그럼에도 불구하고, 많은 후속 연구들이 이러한 파라미터 값에 따른 성능 변화의 분석 없이 기존 연구의 설정 값을 동일하게 사용하고 있다.

본 논문에서는 DCASE 등을 중심으로 최근 연구되고 있는 약지도 음향 이벤트 검출 문제에 대해 살펴보고, 이를 위한 평균-교사 모델 기법에 대해 소개한다. 그리고 공개된 데이터셋을 이용, 파라미터를 변경한 반복 실험을 통해 파라미터 값에 따른 성능을 분석하고자 한다.

II. 약지도 음향 이벤트 검출 문제

최근 시행된 DCASE 2019 및 DCASE 2020 의 Task 4에서 정의된 약지도 음향 이벤트 검출 문제는 다음과 같다. Fig. 1에서 보는 바와 같이, 약지도 음향 이벤트 검출 시스템은 찾고자하는 이벤트 클래스가 특정 오디오 클립에 존재하는지 여부, 그리고 시작/종료 시점을 찾는 것을 목표로 한다. 이는 일반적인 음향 이벤트 문제에서 다루는 바와 동일하다.

약지도 음향 이벤트 검출 문제에서, 학습 데이터는 Fig. 1에서 보는 바와 같이 다양한 종류의 데이터로 이루어져 있다. 학습 데이터 중 일부는 일반적인 음향 이벤트 검출 문제와 마찬가지로 학습을 위한 오디오 클립 데이터, 포함하고 있는 이벤트 클래스의 종류, 그리고 각 이벤트의 시작/종료 시점을 포함하고 있다. 이는 “강하게 표기된 데이터” 라고 명명되었다. 다른 일부는 학습 오디오 클립 데이터와 포함된 이벤트 클래스의 종류만을 가지고 있으며, 이벤트의 시작/종료 시점 데이터는 포함되어 있지 않다. 이를 “약하게 표기된 데이터” 라 부른다. 마지막으로, 일부 데이터는 학습 오디오 클립만 제공되고, 이벤트 클래스의 종류 및 시작/종료 시점 등 모든 레이블이 포함되어 있지 않다. 이는 “미표기 데이터”라 명명되었다.

일반적으로, 학습 데이터를 만드는 데에 있어서 레이블을 추가하는 것이 매우 많은 비용을 필요로 한다. 특히, 이벤트의 시작/종료 시점을 표기하는 것은 단순히 이벤트의 종류만을 표기하는 것보다 더 큰 노력이 필요하다. 따라서, DCASE 2019 Task 4와 같은 경연에서는 매우 적은 수의 강하게/약하게 표기된 데이터와 많은 수의 미표기 데이터를 이용하여 시스템을 학습시키게 된다. 강하게 표기된 데이터의 경우 녹음된 데이터에서 시작/종료 시점을 찾는 대신, 반대로 임의의 시작/종료 시점을 가지도록 이벤트 음향 신호와 배경 신호를 합성하는 경우도 있다.

III. 음향 이벤트 검출 시스템

3.1 음향 이벤트 검출 모델

딥러닝 분야에서 매우 다양한 모델들이 연구되고 있는 만큼, 음향 이벤트 검출을 위하여 종단간(end-to-end) 합성곱 신경망 모델(Convolutional Neural Network, CNN),[12] 합성곱 순환 신경망 모델(Convolutional Recurrent Neural Network, CRNN)[10] 등 다양한 모델들이 시도된 바 있다. 멜-주파수 기반의 특징 값을 입력으로 하는 모델에 대해서 현재 안정적으로 널리 사용되는 것은 CRNN 기반의 모델이다. 최근 트랜스포머[13] 기반의 모델[14]도 좋은 성능을 보인 바 있으나, 아직은 조금 더 검증이 필요할 것으로 판단된다.

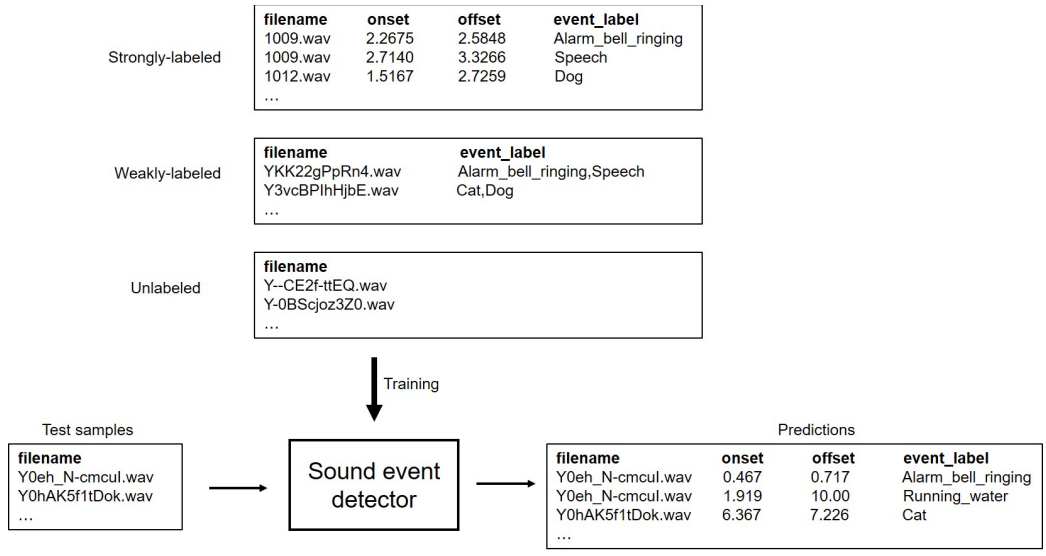

Fig. 1. Problem description for weakly supervised sound event detection.

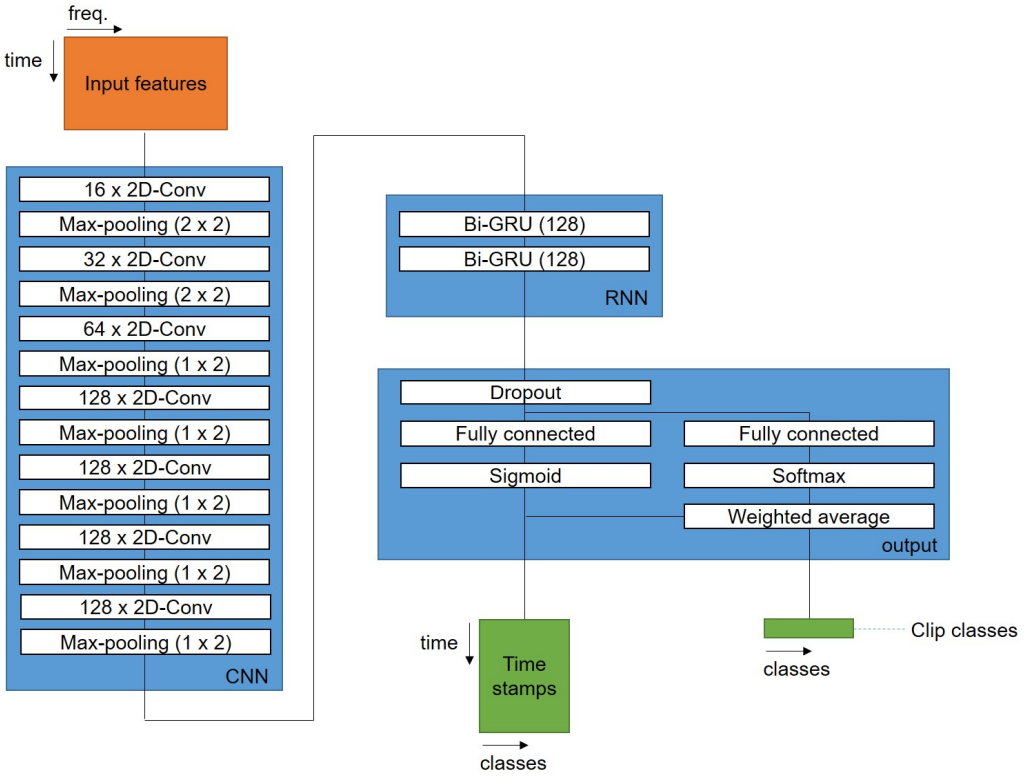

Fig. 2. (Color available online) A diagram of structure for the CRNN classifier.

본 논문에 적용된 CRNN 모델은 Delphin-poulat *et al.*[10]의 모델을 기반으로 하고 있으며, 상세 구조는 Fig. 2와 같다. 모델의 CNN 부분은 기본적으로 2차원 CNN 레이어와 이에 따른 최대값 풀링(max pooling) 레이어로 이루어져 있으며, CNN 레이어를 모두 거쳤을 때 주파수 축의 길이가 1이 되도록 설계되어 있다. Fig. 2에서 보는 바와 같이 매 레이어마다 주파수 축으로 절반의 값을 풀링하도록 구성되어 있으므로, 예를 들어 프레임 별 특징 벡터의 길이가 128인 경우 7개의 CNN 레이어를 사용하면 된다.

풀링 레이어는 각 CNN 레이어의 출력 단에 적용된다. 이 때 주파수 축으로는 매 레이어마다 적용되어 데이터를 절반으로 줄이게 되고, 시간 축으로는 2번만 적용되는 구조이다(즉, 시간 축 방향의 크기는 1/4로 줄어든다).

2차원 CNN 레이어와 풀링 레이어로 이루어진 CNN 모듈의 출력은 순환 신경망(Recurrent Neural Network, RNN) 으로 연결된다. 앞선 CNN 모듈의 최종 출력의 경우 각 프레임 별 출력이 1이 되도록 설계되어 있으므로, RNN 레이어는 프레임 단위의 처리를 수행하게 된다. RNN 레이어의 출력은 완전 연결(fully connected) 레이어로 연결되며, 이 완전 연결 레이어와 출력단의 시그모이드(sigmoid) 활성화 함수는 RNN 출력을 각 클래스의 활성화 확률로 바꾸어 주게 된다. 따라서, 완전연결 레이어의 출력 벡터 크기는 클래스의 개수와 동일하다. 따라서 하나의 오디오 클립에 대해 완전 연결 레이어와 시그모이드 활성화 함수로 얻어지는 출력 데이터의 크기는 (프레임 개수/4) × (클래스 종류의 개수)와 같다. 이는 Fig. 2의 Time stamps 출력단에서 확인할 수 있으며, 매 프레임 당 각 클래스의 존재 확률을 의미한다.

한편, 앞서 언급한 바와 같이 약지도 음향 이벤트 시스템을 학습시키기 위한 데이터는 세 종류의 데이터 – 강하게 표기된 데이터, 약하게 표기된 데이터, 미표기 데이터 – 로 이루어져 있다. 강하게 표기된 데이터는 각 오디오 클립 내에 존재하는 클래스의 종류와 시간 정보를 모두 가지고 있으므로, 이를 이용하여 Time stamps 출력을 위한 정답을 만들 수 있다. 반면, 약하게 표기된 데이터는 클래스의 종류에 대한 정보만을 가지고 있으므로 이러한 정답을 만들

수 없다. 따라서, 약하게 표기된 데이터를 학습하기 위한 출력단이 별도로 필요하다. 약하게 표기된 데이터는 클래스의 존재 여부에 대한 정보만을 가지고 있으므로, 이를 위한 출력단의 데이터 크기는 (1) × (클래스 종류의 개수) 와 같아야 한다. 이와 같은 크기의 압축은 Fig. 2의 가중치 평균(weighted average) 레이어에서 다음과 같이 수행된다.[10]

$$c(j)=\frac{\sum_{n=0}^{N_{out}-1} p(n,j)p_{softmax}(n,j)}{\sum_{n=0}^{N_{out}-1} p_{softmax}(n,j)}, \quad (1)$$

여기서 $c(j)$는 j번째 클래스의 존재 확률을, $p(n,j)$와 $p_{softmax}(n,j)$는 각각 시그모이드 레이어와 소프트맥스(softmax) 레이어의 n번째 프레임, j번째 클래스의 출력을 나타내며, N_{out}은 출력 프레임 개수를 나타낸다.

3.2 평균-교사 모델

평균-교사 모델은 영상 처리 문제에서 레이블이 표기된 데이터와 표기되지 않은 데이터를 학습하기 위해 고안되었다.[8] 평균-교사 모델은 다음과 같은 두 가지 특징을 가지고 있다. 먼저, 평균-교사 모델은 학생 모델과 교사 모델로 이루어져 있으며, 학생 모델과 정답과의 차이에 대한 비용함수(classification cost, 분류 비용) 외에 학생 모델과 교사 모델 사이의 차이에 대한 비용함수(consistency cost, 일관성 비용)를 가진다. 또한, 교사 모델의 계수는 역전파에 의해 갱신되지 않고, 학습 과정에서 다음과 같이 학생 모델 계수의 지수 이동 평균으로 계산된다.[8,10]

$$\theta_{teacher} \leftarrow \alpha\theta_{teacher}+(1-\alpha)\theta_{student}, \quad (2)$$

여기서 $\theta_{teacher}$와 $\theta_{student}$는 각각 교사 모델과 학생 모델의 계수를 의미하고, α는 이동 평균을 조절하는 사용자 설정 파라미터이다.

영상 처리 모델을 위한 평균-교사 모델[8]에서 사용한 분류기는 한 종류의 출력(클래스별 확률) 만을 가지는 반면, 약지도 음향 이벤트 검출 분류기는 두 개

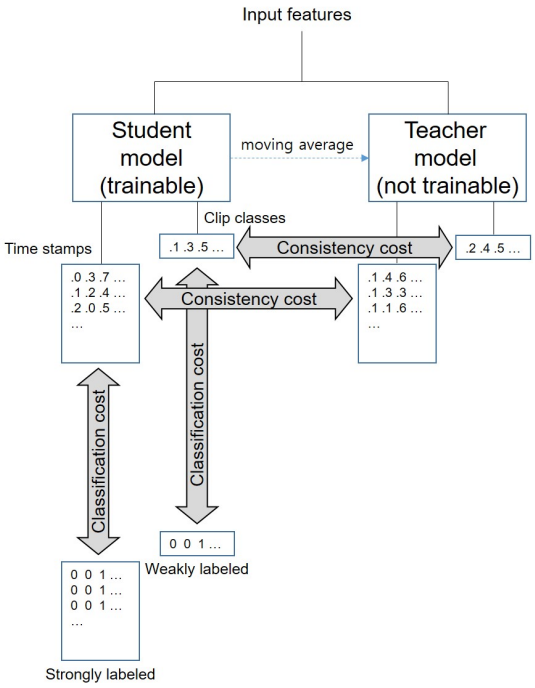

Fig. 3. (Color available online) A schemetic diagram for mean-teacher sound event detection model.

의 출력(Time stamps 출력과 Clip classes 출력, Fig. 2 참고)을 가진다. 따라서 음향 이벤트 검출을 위한 평균-교사 모델[10]은 이에 맞추어 변형될 필요가 있다. 이 모델은 Fig. 3과 같은 형태로 구현될 수 있으며, 강하게 표기된 데이터는 오디오 클립의 특징값과 Time stamps, Clip classes 출력을 위한 정답을, 약하게 표기된 데이터는 특징값과 Clip classes 출력을 위한 정답을, 그리고 미표기 데이터는 특징값만을 제공하게 된다. 이때, 학습을 위한 비용함수는 다음과 같이 정의된다.[8,10]

$$J_{total}(\theta_{student})=J_{cls}(\theta_{student})+\beta J_{con}(\theta_{student}), \quad (3)$$

여기서 $J_{cls}(\theta_{student})$와 $J_{con}(\theta_{student})$는 각각 분류 비용과 일관성 비용을 의미하며, β는 일관성 비용에 대한 가중치로, 사용자 설정 파라미터이다. 분류 비용과 일관성 비용은 각각 Time stamps 출력과 Clip classes 출력에 대한 비용함수를 합산하여 계산하며, 일반적으로 분류 비용으로는 이항 교차 엔트로피를, 일관성 비용으로는 평균 제곱 오차를 사용한다.

3.3 학습률 조절 전략

위와 같은 약지도 음향 이벤트는 약하게 표기된 데이터 혹은 미표기 데이터를 사용하며, 학습 초기에는 교사 모델도 충분히 학습되지 않았기 때문에 이와 같은 데이터들은 큰 학습 오차를 만들어낼 수 있다. 이러한 문제를 해결하기 위해 학습 초기에는 낮은 학습률을 적용하였다가 점차 이를 증가시키는 전략을 사용한다. 이를 램프-업(ramp-up)이라 하며, 평균-교사 모델에서는 다음과 같은 지수적 램프-업을 널리 사용한다.[10]

$$\mu(n)=\begin{cases} e^{-\gamma(n)}\mu_{max}, & \text{if } n \le N_{rampup} \\ \mu_{max}, & \text{if } n > N_{rampup} \end{cases}, \tag{4}$$

여기서 $\mu(n)$은 n번째 에포크의 학습률을 의미하며, μ_{max}는 최대 학습률, N_{rampup}은 램프-업 구간의 길이(에포크 수), 그리고 $\gamma(n)$은 램프-업 계수로 다음과 같이 계산된다.[10]

$$\gamma(n)=5\left(1-\frac{n}{N_{rampup}}\right)^2. \tag{5}$$

위에서 살펴본 바와 같이, 평균-교사 모델 기반의 약지도 음향 이벤트 검출 시스템은 교사 모델의 이동 평균 조절 파라미터 α, 일관성 비용함수에 대한 가중치 β, 램프-업 구간의 길이 N_{rampup}, 그리고 최대 학습률 μ_{max} 등의 사용자 설정 파라미터를 가지며, 이러한 파라미터는 모두 학습 및 분류 성능에 영향을 미치게 된다. 본 논문에서는, 실제 녹음된 음향 이벤트 데이터베이스에 대한 시뮬레이션을 통하여 각 파라미터가 성능에 미치는 영향을 분석하고자 한다.

IV. 시뮬레이션 및 결과 분석

4.1 시뮬레이션 환경

위와 같이 구현한 평균-교사 모델 기반의 약지도 음향 이벤트 시스템의 파라미터에 따른 성능을 분석하기 위하여, DCASE 2020 Task 4[4] 의 개발 데이터를 이용하여 성능을 분석하였다. 학습 데이터는 2045개의 강하게 표기된 데이터와 1578 개의 약하게 표기된 데이터, 그리고 14412개의 미표기 데이터로 구성되었다. 각 오디오 클립은 10 s 길이를 가지고 있으며, 16 kHz의 샘플링 주파수를 가지도록 처리되었다. 각 오디오 클립은 2048 샘플 길이의 프레임(255 샘플의 겹침) 당 128 개의 특징값으로 변환되었다.

본 시뮬레이션에서는, 1) 특징 값의 종류, 2) 이동 평균 파라미터 α, 3) 일관성 비용함수의 가중치 β, 4) 램프-업 길이 N_{rampup}, 5) 최대 학습률 $\mu_{\max}$을 변화시켜가며 음향 이벤트 검출 시스템 성능의 변화를 관찰하고자 한다. 하나의 파라미터를 변화시키는 동안 다른 파라미터는 하나의 값으로 고정시켰다. 고정된 파라미터 값은 Delphin-Poulat 의 모델[10]에 준하여 결정되었으며, 구체적인 값은 다음과 같다. 특징 값으로는 128 개의 주파수 빈을 가지는 멜-스펙트럼을 사용하였고, 이동 평균 파라미터는 0.999, 일관성 비용함수의 가중치는 2, 램프-업 길이는 50 에포크, 그리고 최대 학습률은 0.001의 값을 사용하였다.

학습은 총 200 에포크 동안 진행되었고, 그 중 검증 데이터에 대해 가장 작은 비용함수를 가지는 모델을 저장하여 성능 평가에 사용하였다. 배치 크기는 24로 설정되었고, 구성된 모델은 Adam[15] 기법을 이용하여 학습되었다. 검출기의 출력 확률값은 0.5의 문턱값을 기준으로, 문턱값보다 큰 경우 이벤트가 존재하는 것으로, 그렇지 않은 경우 이벤트가 존재하지 않는 것으로 판별하였다.

결과에 대한 성능을 평가하기 위한 지표로, 이벤트-기반 F1-score와 세그먼트-기반 F1-score가 사용되었다. F1-score는 정확도 지표 중 하나로, 다음과 같은 정밀도 P와 재현율 R의 기하 평균으로 계산된다.

$$P=\frac{n_{TP}}{n_{TP}+n_{FP}}, \tag{6}$$

$$R=\frac{n_{TP}}{n_{TP}+n_{FN}}, \tag{7}$$

$$F_1=\frac{2PR}{P+R}, \tag{8}$$

여기서 n_{TP}는 이벤트가 존재하는 것을 정확히 맞춘 개수를, n_{FP}는 이벤트를 검출하였지만 실제로는 이벤트가 없는 경우(오탐지)의 개수를, 그리고 n_{FN}는 실제 이벤트가 존재하지만 검출기가 놓친 개수를 의미한다. 세그먼트-기반 지표는 클립을 1초 구간으로 나누어 평가하며, 각 1 s 구간마다 추정치와 정답을 비교하여 성능을 평가한다. 이벤트-기반 지표는 각 이벤트 구간 별로 추정치와 정답을 비교하여 성능을 평가하며, 이 때 이벤트의 시작 시점과 종료 시점을 모두 맞춘 경우 이를 정확한 답으로 판정한다. 이 때 시작 시점은 0.2 s의 오차 범위 내에 있는 경우 이를 정답으로 인정하고, 종료 시점은 0.2 s(혹은 이벤트 길이의 20 %)의 오차 내에 있는 경우 정답으로 인정한다. 이는 DCASE 2019 Task 4 및 DCASE 2020 Task 4의 기준과 동일하다.

4.2 시뮬레이션 결과

첫 번째로, 특징 값의 종류에 따른 성능의 변화를 관찰하였다. 적용된 특징 값은 음향 신호 분류 및 음향 이벤트 검출 등의 과업에서 널리 사용되는 멜-스펙트럼, 로그-멜 스펙트럼, 그리고 감마톤 스펙트럼을 적용하여 비교하였다. 각 특징 벡터의 크기는 128로 동일하게 설정되었으며, 프레임 길이 또한 앞서 기술한 바와 동일하다.

Table 1은 각 특징 별 성능을 비교하여 나타내고 있다. Micro-averaged F1-score는 클래스와 관계없이 모든 정답을 누적하여 계산한 평균 성능이고, macro-averaged F1-score는 각 클래스 별 F1-score를 먼저 계산한 후, 이에 대한 평균을 계산한 지표이다. 각 클래스 별로 정답의 개수가 다르기 때문에, 정답이 많은 클래스의 성능이 높은 경우 micro-averaged F1-score의 지표가 더 높을 것이고, 모든 클래스의 성능이 고르게 나타난 경우 macro-averaged F1-score의 성능이 더 좋을 수 있다.

성능 결과를 살펴보면, 모든 지표에 대하여 멜-스펙트럼의 성능이 가장 좋고, 그 다음이 감마톤 스펙트럼, 그리고 로그-멜 스펙트럼의 성능이 가장 낮게 나타났다. 멜-스펙트럼과 감마톤 스펙트럼의 성능 차이는 1 %~2 % 정도로 크지 않았으나, 로그-멜-스펙트럼은 4 %~8 % 정도의 큰 성능 저하를 보였다.

다음으로, 이동 평균 파라미터 α에 따른 성능을 분석하였다. Eq. (2)에서 보는 바와 같이, 이동 평균 파라미터가 클수록 새로운 학생 모델 계수에 대한 가중치가 작다. 즉, 교사 모델의 계수가 느리게 변하면서 더욱 긴 데이터에 대한 평균을 계산하는 셈이 된다.

이동 평균 파라미터에 대한 성능은 Table 2에 정리되어 있다. 이를 통해 Delphin-poulat *et al*.[10] 여러 모델에서 널리 사용되는 0.999의 값보다는 0.99의 값이 더 좋은 성능을 보이는 것을 확인할 수 있었다. $\alpha = 0.9999$인 경우 미시-평균 이벤트 기반 성능이 높고, 거시-평균 성능이 나쁜 것을 확인할 수 있는데, 이는 이벤트 개수가 많은 클래스에서 더 좋은 성능이 나온 것을 의미한다. 테스트 이벤트 개수가 많은 클래스는 대부분 훈련 데이터 개수도 많기 때문에, 데이터가 많은 클래스에 치중하여 학습이 된 것을 알 수 있다.

Table 3은 일관성 비용 함수에 대한 가중치 β에 따른 성능 비교를 나타내고 있다. 평균-교사 모델 기반의 약지도 음향 이벤트 검출기에서 널리 사용되는 값 ($\beta = 2$)과 달리 4~16 사이의 값이 일반적으로 좋은 성능을 보였다. 다만 모든 지표에서 좋은 성능을 보이는 β 값은 존재하지 않고, 지표의 종류와 평균

Table 1. Averaged results with various features. MelSpec, LogMel, and GAM denote the results of Mel-spectrum, log-Mel-spectrum, and gammatone spectrum, resepectively.

Feat.	Micro-avg. F1-score (event)	Macro-avg. F1-score (event)	Micro-avg. F1-score (segment)	Macro-avg. F1-score (segment)
MelSpec	**34.82**	**34.48**	**74.89**	**68.93**
LogMel	30.27	29.88	67.74	60.69
GAM	32.15	33.09	73.96	67.81

Table 2. Averaged results with various moving average factors of the teacher model.

α	Micro-avg. F1-score (event)	Macro-avg. F1-score (event)	Micro-avg. F1-score (segment)	Macro-avg. F1-score (segment)
0.9	32.99	33.60	74.35	67.30
0.99	37.78	**37.72**	**75.96**	**70.45**
0.999	34.82	34.48	74.89	68.93
0.9999	**41.14**	31.77	73.65	67.05

Table 3. Performance comparisons with various weights of the consistency cost.

β	Micro-avg. F1-score (event)	Macro-avg. F1-score (event)	Micro-avg. F1-score (segment)	Macro-avg. F1-score (segment)
2	34.82	34.48	74.89	68.93
4	39.10	36.84	**75.73**	**70.56**
8	39.88	**38.28**	74.43	68.65
16	40.61	36.84	74.43	68.05
32	**41.38**	34.92	71.73	62.49
64	37.47	28.51	65.29	50.89

Table 4. Comparison results with different ramp-up lengths.

N_{rampup}	Micro-avg. F1-score (event)	Macro-avg. F1-score (event)	Micro-avg. F1-score (segment)	Macro-avg. F1-score (segment)
10	35.54	34.94	73.48	66.94
30	**35.60**	**35.47**	73.86	68.25
50	34.82	34.48	**74.89**	**68.93**

을 구하는 방법에 따라 최적의 β 값이 달라졌다. $\beta=32$의 경우는 미시-평균 이벤트 기반 성능은 성능이 높았으나, 나머지 3개의 지표에서 낮은 성능을 보였다.

Table 4는 램프-업 길이 N_{rampup}에 따른 성능을 비교하였다. $N_{rampup}=30$일 때에는 이벤트-기반 성능이 좋고, $N_{rampup}=50$일 때에는 세그먼트-기반성능이 좋은 경향을 보였다. 다만, 다른 파라미터에 비해 램프-업 길이의 차이는 성능에 큰 영향을 주지 않는 것을 볼 수 있었다.

Table 5는 최대 학습율 μ_{max}에 따른 성능의 변화를 보여주고 있다. Delphin-poulat *et al.*[10]의 모델에서 사용한 값은 0.001이었으나, 이보다는 작은 학습율(0.0001 ~ 0.0005)에서 더 좋은 성능을 보였다. 0.001의 최대 학습율은 미시-평균 세그먼트-기반 지표에서 가장 좋은 성능을 보였으나, 나머지 지표에서 다소 저하된 성능을 보였다. 가장 좋은 이벤트-기반 성능은 0.0001의 학습율에서, 거시-평균 세그먼트-기반 성능은 0.0002의 학습율에서 볼 수 있었으나, 둘 사이의

Table 5. Comparison results with various maximum learning rates.

μ_{max}	Micro-avg. F1-score (event)	Macro-avg. F1-score (event)	Micro-avg. F1-score (segment)	Macro-avg. F1-score (segment)
0.00005	36.38	35.37	71.82	67.22
0.0001	**38.38**	**38.30**	74.53	69.24
0.0002	37.91	37.87	74.56	**69.40**
0.0005	38.33	37.06	73.24	68.31
0.001	34.82	34.48	**74.89**	68.93
0.005	28.28	24.47	68.07	58.50
0.01	27.23	24.58	67.72	59.25

성능 차이는 크지 않았다.

V. 결 론

본 논문에서는, 강하게 및 약하게 표기된 데이터, 그리고 미표기 데이터를 이용하여 학습할 수 있는 약지도 음향 이벤트 검출 시스템을 구현하고, 파라미터 설정에 따른 성능의 차이를 분석하였다. 본 논문에서 구현된 약지도 음향 이벤트 검출 시스템은 최근 DCASE 등에서 널리 사용되고 있는 평균-교사 모델이 적용된 합성곱 순환 신경망 모델의 구조를 기반으로 하였다. 평균-교사 모델은 표기 및 미표기 데이터를 이용하여 모델 계수를 학습하는 학생 모델과, 이에 대한 이동 평균 값을 계수로 가지는 교사 모델로 이루어진다. 학생 모델은 추정치와 정답에 대한 오차, 그리고 교사 모델과의 추정치에 대한 오차에 대한 역전파를 통해 계수를 학습하며, 교사 모델은 역전파에 의해 계수를 학습하지 않는 대신 학생 모델의 이동 평균으로 얻어진다.

평균-교사 모델은 약지도 학습을 위해 추가적인 구조를 가지고 있기 때문에, 그만큼 사용자가 설정할 수 있는 파라미터의 종류 더 많다. 이는 결국 파라미터의 설정 값에 따라 성능이 변할 수 있다는 것을 의미하며, 이에 대한 영향을 분석하기 위하여 DCASE 2020 Task 4 데이터를 이용하여 시뮬레이션을 진행하였다. 시뮬레이션 결과, 멜-스펙트럼과 감마톤 스펙트럼의 성능이 비교적 높게 나타났으며, 이동 평균 파라미터의 경우 일반적으로 많이 사용하는 0.999 보다 0.99의 값에서 더 좋은 성능이 나타났고, 비교적

성능에 미치는 영향이 큰 편이었다. 일관성 비용함수의 가중치는 4 ~ 16 사이의 값에서 좋은 성능을 보였고, 램프-업 길이는 성능에 미치는 영향이 크지 않았다. 최대 학습율 또한 0.0001 ~ 0.0002에서 좋은 성능을 보였는데, 이는 널리 사용되는 0.001의 값과는 다소 달랐다.

감사의 글

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2020R1I1A3069162).

References

1. D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” IEEE Signal Process. Mag. **32**, 16-34 (2015).
2. E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, “Polyphonic sound event detection using multi label deep neural networks,” Proc. IJCNN. 1-7 (2015).
3. S. Lee and H.-S.Pang, “Feature extraction based on the non-negative matrix factorization of convolutional neural networks for monitoring domestic activity with acoustic signals,” IEEE Access, **8**, 122384-122395 (2020).
4. N. Trupault, R. Serizel, A. P. Shah, and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” Proc. 2019 DCASE Workshop, 253-257 (2019).
5. P. Bachman, O. Alsharif, and D. Precup, “Learning with pseudo-ensembles,” Advances in Neural Information Processing Systems, **27**, 3365-3373 (2014).
6. M. Sajjadi, M. Javanmardi, and T. Tasdizen, “Regularization with stochastic transformations and perturbations for deep semi-supervised learning,” Advances in Neural Information Processing Systems, **30**, 1163-1171 (2016).
7. S. Laine and T. Alia, “Temporal ensembling for semi-supervised learning,” arXiv preprint arXiv:1610.02242 (2016).
8. A. Tarvainen and H. Valpola, “Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results,” Advances in Neural Information Processing Systems, **31**, 1195-1204 (2017).
9. L. JiaKai and P. Shanghai, “Mean teacher convolution system for dcase 2018 task 4,” DCASE. 2018 Challenge Tech. Rep., 2018.
10. L. Delphin-Poulat and C. Plapous, “Mean teacher with data augmentation for dcase 2019 task 4,” Orange Labs Lannion Tech. Rep., 2019.
11. J. Yan and Y. Song, “Weakly labeled sound event detection with residual crnn using semi-supervised method,” DCASE 2019 Challenge Tech. Rep., 2019.
12. S. Lee, M. Kim, and Y. Jeong, “A study on the waveform-based end-to-end deep convolutional neural network for weakly supervised sound event detection” (in Korean), J. Acoust. Soc. Kr. **39**, 24-31 (2020).
13. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems, **31**, 5998-6008 (2017).
14. K Miyazaki, T. Komatsu, and T. Hayashi, “Convolution augmented transformer for semi-supervised sound event detection,” Proc. 2019 DCASE Workshop, 100-104 (2019).
15. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980 (2014).

저자 약력

▸ 이 석 진 (Seokjin Lee)

2006년 8월 : 서울대학교 전기컴퓨터공학부 학사
2008년 8월 : 서울대학교 전기컴퓨터공학부 석사
2012년 2월 : 서울대학교 전기컴퓨터공학부 박사
2012년 3월 : ㈜LG전자 CTO연구소 선임연구원
2014년 3월 : 경기대학교 전자공학과 조교수
2018년 3월 : 경북대학교 전자공학부 조교수
2020년 10월 ~ 현재 : 경북대학교 전자공학부 부교수