Journal of KIIT. Vol. 20, No. 6, pp. 93-100, Jun. 30, 2022. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2022.20.6.93

음향 이벤트 탐지를 위한 멀티-스케일 특징 기반 트랜스포머 모델

김수종*, 정용주**

A Transformer Model based on Multi-scale Features to Improve the Performance of Sound Event Detection

Soo-Jong Kim*, Yong-Joo Chung**

요 약

본 연구에서는 소리 탐지를 위해 멀티-스케일 특징을 활용하는 방법을 제안하였다. 이를 위해 소리 신호의 시계열 상관관계 모델링에 있어서 기존의 RNN(Recurrent Neural Network)에 비해서 우수한 성능을 보인 트랜스포머-인코더 기반의 심층신경망 구조에 특징-피라미드 기법을 적용하였다. 제안된 방법인 멀티-스케일 특징을 사용함으로써, 기존의 심층신경망 모델보다 클래스별 다양한 소리 신호의 길이 변화에 더욱 강인해질 수 있다. 본 연구에서 제안된 방법을 DCASE 2019 Task 4 데이터셋에 대해 실험하고 평가하였으며, 멀티-스케일 특징을 사용하지 않은 기존의 심층신경망 모델에 비해 상대적 개선도가 5.4% 더 우수함을 확인할 수 있었다.

Abstract

We propose a method that utilizes multi-scale features for sound event detection. We employed a feature-pyramid component in a deep neural network architecture based on the transformer encoder that is used to model the time correlation of sound signals owing to its superiority over conventional recurrent neural networks. The proposed method is motivated by the idea that the multi-scale features will make the network more robust against the dynamic duration of the sound signals depending on their classes. We conducted experiments using the DCASE 2019 Task 4 dataset to evaluate the performance of the proposed method. The experimental results show that the proposed method outperforms the baseline neural network without multi-scale features.

Keywords
sound event detection, transformer encoder, feature-pyramid, convolutional neural network, attention model

* 계명대학교 전자공학과 석사과정
- ORCID: https://orcid.org/0000-0001-7492-1134
** 계명대학교 전자공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-0060-1178

· Received: Mar. 21, 2022, Revised: May 16, 2022 Accepted: May 19, 2022
· Corresponding Author: Yong-Joo Chung
Dept. of Electronics Engineering, Keimyung University, 704-701
Shindang-dong, Dalseo-gu, Daegu-si, 1000, Republic of Korea,
Tel.: +82-53-580-5925, Email: yjjung@kmu.ac.kr

Ⅰ. 서 론

일상생활에서 소리는 보통 주변 상황에 대한 중요한 정보를 포함하고 있으며, 소리 신호에서 정보를 자동으로 추출하기 위한 많은 연구와 노력이 머신러닝 패러다임 속에서 이루어지고 있다. 음향 이벤트 검출(SED, Sound Event Detection)은 2013년부터 2021년까지 진행된 DCASE(Detection and Classification of Acoustic Scenes and Events) Challenge에서 주된 주제 중 하나로, 음향 신호의 종류를 분류하고 음향 이벤트의 시작점과 끝점을 찾는 것을 목표로 한다. SED의 적용 분야는 다양하며 감시나 도시 소음 분석, 멀티미디어 컨텐츠로부터의 정보 탐색, 헬스케어 모니터링 및 새소리 탐지 등의 분야를 포함한다[1]-[3].

심층신경망 구조(DNN, Deep Neural Networks)가 컴퓨터 비전[4]이나 음성인식[5], 기계번역[6] 등에서 가장 우수한 성능(State-of-the-art)을 보였기 때문에 현재의 SED 연구는 주로 DNN 기반의 접근법에 초점을 맞추고 있다[7]-[9]. 그중에서 CNN (Convolutional Neural Network)와 RNN(Recurrent Neural Network)을 합친 구조인 CRNN(Convolutional Recurrent Neural Networks)은 SED에서 가장 만족스러운 분류 성능을 보여[7] 현재 SED에서 대표적인 심층신경망 구조로 간주 되고 있다.

SED에서 중요한 문제 중 하나는 음향 신호의 시간적 상관관계(Temporal correlations)를 모델링하는 것이다. 신호 분석의 관점에서 모델링할 수 있는 짧은 시간적 상관관계에 비해 긴 시간적 상관관계는 다소 모델링을 하기가 어렵다. 이러한 긴 상관관계에 관한 순서열 데이터에서 정보를 추출하는 모델로 RNN이 널리 사용되고 있지만 RNN은 모델링할 수 있는 상관관계의 길이(Correlation length)가 한정적이기 때문에 성능 면에서 만족스럽지 못한 부분이 있었다. 이 문제를 극복하기 위해 멀티-스케일(Multi-scale) 특징을 RNN의 입력으로 사용하는 몇몇 소리 신호 분류 연구가 진행되었다[10]-[12].

특징-피라미드 네트워크(FPN, Feature-Pyramid Network)는 컴퓨터 비전 영역에서 객체 인식(Object detection) 문제를 위해 사용되고 있는 모델이다[13]. FPN은 멀티-스케일 특징을 명시적으로 추출하는 대신 CNN의 저해상도 층의 특징맵과 고해상도 층의 특징맵을 결합함으로써 멀티-스케일 특징을 생성하고, 이를 통해서 스케일 변동을 상쇄함으로써 객체 감지 문제를 개선했으며 다양한 스케일의 이미지에서 객체를 인식하는데 좋은 성과를 거두었다. 컴퓨터 비전에서의 FPN의 성공에 영향을 받아, 음향 탐지 분야에서도 SED의 성능을 향상시키기 위해 특징-피라미드 구성 요소를 갖춘 CRNN 구조가 제안되었다[14].

최근 연구에서, 트랜스포머(Transforemer) 모델은 언어 순서열 데이터를 모델링하는데 있어 RNN보다 우수하다고 여겨지고 있으며[15], 음성/오디오 등의 처리 분야에서도 널리 쓰이고 있다[16]-[18]. [17]에서는 트랜스포머를 SED에 적용하였으며, 그 성능에 있어서 Baseline CRNN보다 뛰어남을 보여주었다.

본 연구에서는 트랜스포머 인코더에 기반한 심층신경망에 FPN에 근거한 멀티-스케일 특징을 SED에 적용할 것을 제안하였다. 멀티-스케일 특징은 다양한 시간 분해능에 대한 정보를 담고 있어 같은 종류의 소리라도 지속시간이 유동적인 음향신호에 대해 더욱 강인한 특성을 가진다. 해당 특징은 시간 상관관계에 대한 정보 처리 시 트랜스포머 인코더의 기능 향상에 기여할 것으로 예상되므로, 제안한 방법이 기존의 방식에 비해서 SED에서 더 나은 성능을 보여 줄 것으로 기대한다. 해당 구간을 추가함으로써 428,544(+396,544 * Encoder's layer)만큼의 파라미터가 추가되지만 GLU 활성화함수를 사용하여 Overfitting 및 Underfitting에 대한 문제를 보완하였다. 우리가 아는 한, 트랜스포머 인코더 기반 심층신경망에서 멀티-스케일 특징을 사용한 것은 본연구가 처음이다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에 사용된 트랜스포머 인코더에 대한 설명을, 3장에서는 제안된 심층신경망의 구조를 설명한다. 4장에서는 실험에 사용한 데이터베이스를 소개하고 본 연구에서 실험한 다양한 결과를 제시한다. 마지막으로 5장에서는 결론을 맺는다.

Ⅱ. 트랜스포머 인코더

트랜스포머 인코더(Encoder) 구조는 인코더 블록

(Block)의 겹침으로 구성되며 상세 구조는 그림 1과 같다. 각각의 인코더 블록은 멀티-헤드 어텐션(Multi-head attention) 층과 포지션-와이즈 피드포워드(Position-wise feed-forward) 층으로
나뉘며 각 층은 드롭아웃(Dropout)을 적용한 후 잔차 연결(Residual connection)과 레이어 정규화(Layer normalization)를 적용하는 구조로 되어있다.

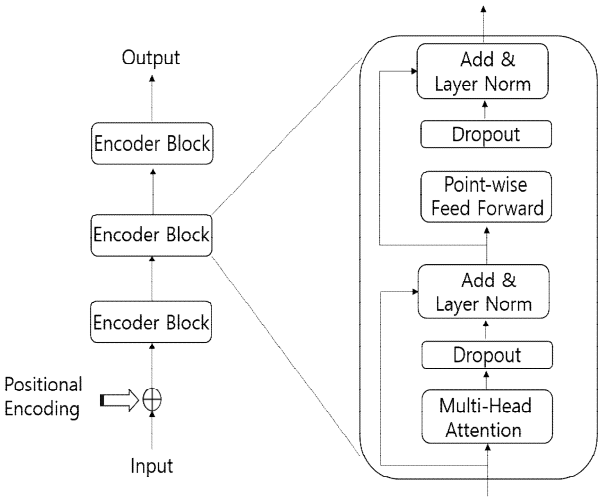

그림 1. 트랜스포머 인코더의 구조
Fig. 1. Structure of the transformer encoder

2.1 포지셔널 인코딩

트랜스포머 인코더에는 순환 특성(Recurrence)이 포함되어 있지 않으므로 입력 데이터의 순서 정보를 모델에 추가해야 한다. 이를 위해 데이터의 위치 순서에 해당하는 일부 정보를 인위적으로 생성하여 입력 데이터에 추가해야 하며 이를 포지셔널 인코딩(Positional encoding)이라고 한다. 포지셔널 인코딩은 식 (1)과 같다. [15]

$$PE_{(pos,\ 2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{1}$$
$$PE_{(pos,\ 2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

pos는 입력 데이터 순서열(Sequence)에서의 위치를 나타내며, i는 순서열 데이터의 차원을 나타내며, 그리고 d_{model}은 어텐션 유닛의 개수를 나타낸다.

2.2 멀티 헤드 어텐션

트랜스포머 인코더에 사용되는 어텐션 메커니즘을 스케일 내적 어텐션(Scaled dot-product attention)이라고 하며 다음과 같이 정의한다.

$$\text{Attention(Q,K,V)=softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

식 (2)에서 어텐션 출력은 밸류(Value) 값 V의 가중치 합계로 계산된다. 값에 할당된 가중치는 쿼리(Query) 값 Q와 키(Key) 값 K의 내적으로 계산된다. d_k는 키의 차원이며 T는 행렬의 전치(Transpose)를 의미한다. Q=K=V일 때, 사용된 트랜스포머 인코더의 어텐션 기법을 셀프 어텐션(Self-attention)이라고 한다.

식 (2)의 어텐션 메커니즘을 여러 번 수행하는 방식을 멀티 헤드 어텐션이라 한다. 멀티 헤드 어텐션은 쿼리 값, 키 값, 밸류 값을 선형 변환으로 투영(Project)한 후 어텐션 기법을 수행한다. 이를 통해서 얻은 h개의 서로 다른 어텐션 출력들은 연결(Concatenate)되고 투영되어서 최종 출력이 된다. 멀티 헤드 어텐션은 다음과 같이 정의된다[15].

$$\text{Multihead(Q,K,V)=Concat}\left(H_1, H_2, \cdots, H_h\right) W^O$$
$$H_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{3}$$

여기서 $W_i^Q \in \mathcal{R}^{d_{model} \times d_k}$, $W_i^K \in \mathcal{R}^{d_{model} \times d_k}$, $W_i^V \in \mathcal{R}^{d_{model} \times d_v}$, $W_i^O \in \mathcal{R}^{hd_v \times d_{model}}$이며 $d_k = d_v = d_{model}/h$ 이다.

2.3 포지션-와이즈 피드 포워드 신경망

포지션-와이즈 피드 포워드 신경망(Position-wise feed-forward networks)은 멀티-헤드 어텐션 층 다음에 적용되는 층으로 각 위치에 개별적으로 동일하게 완전 연결 신경망(Fully connected networks)이 적용된다. 해당 네트워크는 2개의 선형 층으로 구성되고 다음과 같이 정의된다[15].

$$\text{FFN}(X(t)) = \max\left(0,\ X(t)\,W_1 + b_1\right)W_2 + b_2 \tag{4}$$

$X(t)$는 입력 순서열 데이터 X의 t번째 프레임을 의미하고 $W_1 \in \mathcal{R}^{d_{model} \times d_{ff}}$, $W_2 \in \mathcal{R}^{d_{ff} \times d_{model}}$, $b_1 \in \mathcal{R}^{d_{ff}}$, $b_2 \in \mathcal{R}^{d_{model}}$ 이다.

III. 트랜스포머 네트워크 구조

3.1 네트워크 구조

본 논문에서 사용한 트랜스포머 모델의 구조는 CRNN을 기반으로 한 특징-피라미드 네트워크를 사용했다[14]. 제안된 모델은 베이스라인 트랜스포머(Baseline transformer), 멀티-스케일 특징 추출(Multi-scale feature extraction) 그리고 분류 층(Classification layer)의 3가지로 이루어져 있으며 구조는 그림 2에 자세히 나와 있다.

베이스라인 트랜스포머는 7개의 콘볼루션 블록과 하나의 트랜스포머 인코더로 구성되어 있다. 입력으로는 전처리를 통해 만들어진 2차원의 로그-멜 스펙트럼 특징맵(Feature map)이 사용된다. 콘볼루션 블록에서는 3x3의 콘볼루션 필터가 적용되었으며 배치정규화(Batch normalization)가 사용되었다. 배치정규화 이후 GLU(Gated Linear Unit)가 활성화 함수로 적용되고, 훈련 중 과적합을 줄이기 위해 드롭아웃(Dropout)이 적용된다. 콘볼루션 블록의 마지막 출력은 트랜스포머 인코더의 입력으로 사용된다.

입력 데이터의 시간과 주파수 영역에서의 차원을 위해 주파수와 시간 영역에 {2×2, 1×2, 1×1}의 평균 풀링(Average pooling)이 선택적으로 적용되었다. 이를 통해서 주파수 영역은 하나의 프레임으로 축소되었지만, 시간 영역은 트랜스포머 인코더에서 소리 신호의 시간 상관 정보를 사용하기 위해 62프레임으로 축소되었다. 본 논문의 트랜스포머 인코더는 128의 어텐션 유닛(d_{model}=128), 16개의 병렬 헤드(h=16), 512개의 내부 포지션-와이즈 피드 포워드 신경망(d_{ff}=512)으로 구성된다.

멀티-스케일 특징 추출 구간에서는 콘볼루션 블록과 결합한 트랜스포머 인코더 2개가 추가로 사용되었다. 해당 구간에서는 두 번째 트랜스포머 인코더의 출력이 업샘플링(Upsampling) 되고 첫 번째 트랜스포머 인코더의 출력과 Concatenate 방식으로 합쳐진다. 그 후 합쳐진 특징맵에 1x1 콘볼루션이 적용되어 데이터를 더 부드럽게 하고 차원을 감소시킨다. 이 결과물은 다시 한번 더 업샘플링 되고 베이스라인의 트랜스포머 인코더의 출력과 연결 후 1x1 콘볼루션이 적용된다.

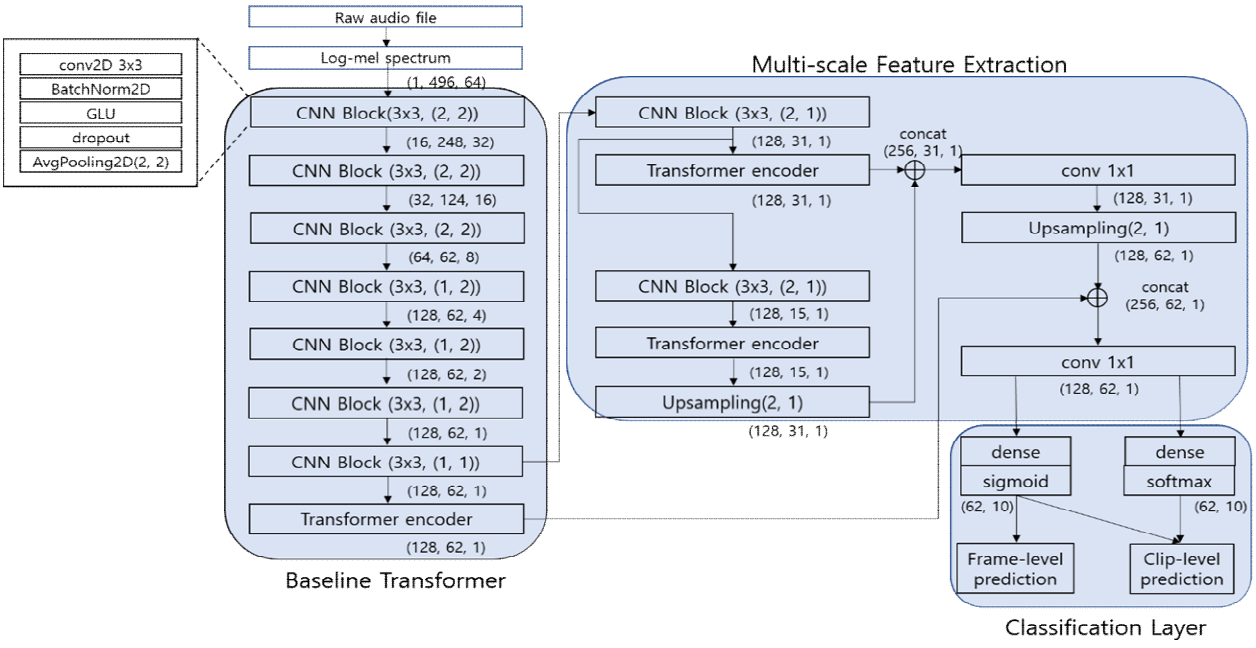

그림 2. 제안된 트랜스포머 네트워크 모델
Fig. 2. Proposed transformer network model

이렇게 얻은 멀티-스케일 특징은 분류 층의 입력으로 사용된다. 2,045개의 합성된 오디오 클립으로 구성된 강한 레이블 데이터, 1.548개의 실제 소리를 녹음한 오디오 클립으로 구성된 약한 레이블 데이터, 그리고 14,412개의 실제 소리를 녹음한 레이블이 없는 데이터의 3가지로 분류된다.

실제 소리를 녹음한 오디오 클립은 구글의 오픈 소스인 Audioset[19]에서 가져왔으며, 합성된 오디오 클립은 FSD dataset[20]의 끗한 음향 신호와 SINS dataset[21]의 잡음 섞인 음향 신호를 합성한 오디오 클립이다. 오디오 클립의 길이는 10초이며 분류하고자 하는 소리의 종류는 10가지로 가정에서 흔히 발생하는 소리 이벤트로 구성되어 있다. 제안된 방법의 성능 테스트는 DCASE 2019 Challenge task 4의 공식 테스트 데이터 세트인 검증 데이터(Validation data) (1168 clips)와 평가 데이터(Evaluation data) (692 clips)를 사용하여 진행했다.

사용자 설정 파라미터로, 배치의 크기는 128개이며, Optimizer는 Adam[22]을 사용하였고 Learning rate는 0.001로 설정하였으며 Pytorch의 StepLR scheduler를 학습 Iteration동안 학습률을 줄이기 위한 방법으로 사용했다. 배치 당 입력 데이터의 비율은 강한 레이블, 약한 레이블, 레이블이 없는 학습 데이터를 각각 1:1:2의 비율로 설정했다. 학습은 3100번의 Mini-batch iteration steps만큼 진행되었다.

3.2 실험 결과

트랜스포머 모델의 학습은 약한 레이블과 강한 레이블 데이터로만 진행하였다. 트랜스포머 모델의 인코더 블록의 수(NEB, Number of Encoder Blocks)가 성능에 미치는 영향을 조사하기 위해 블록의 수를 1에서 3까지 변경하면서 실험했으며 테스트 결과의 신뢰성을 확보하기 위해 각 조건에서 실험(훈련/테스트)을 10회 수행한 후 결과들의 평균을 통해서 F-Score와 Error Rate(ER)을 구하였다.

표 1에서는 멀티-스케일 특징의 사용과 후처리 과정(Post processing)의 유무에 따른 검증 데이터에 대한 이벤트 기반 SED 결과를 볼 수 있다. 후처리 과정은 일정한 임계값(0.5)에 따라 네트워크의 출력을 이진화한 후, 중간값 필터링(Median filtering)을 적용했다. 중간값 필터링을 적용하는 프레임의 길이는 소리 클래스에 따라 달라지며 이러한 방식은 길이가 고정된 필터를 사용하는 것보다 모델의 성능을 더 향상시키는 것으로 확인되었다[18]. 표 1의 결과에서는 멀티-스케일 트랜스포머가 후처리 적용 여부에 관계없이 베이스라인 트랜스포머의 성능을 능가하지만, 후처리를 적용했을 경우에는 전체적인 성능이 훨씬 우수하다는 것을 알 수 있다. 인코더 블록의 수가 다양하게 변해도 안정적인 F-score를 나타내는 멀티-스케일 트랜스포머에 비해 베이스라인 트랜스포머는 인코더 블록의 수가 변화함에 따라 성능이 크게 변동한다. 멀티-스케일 특징을 사용하는 것이 트랜스포머 모델을 더 견고하게 하고 성능을 향상시킨다고 할 수 있다. 후처리를 적용하지 않았을 때 평균 F-score가 베이스라인 트랜스포머에서는 26.51%인 반면 멀티-스케일 트랜스포머는 27.79%를 달성하였으며, 4.8%의 상대적인 F-score 향상을 얻을 수 있었다.

표 1. 베이스라인 트랜스포머 모델과 멀티-스케일 트랜스포머 모델 간 검증 데이터에 대한 이벤트 기반 SED 결과 비교
Table 1. Comparison of event-based SED results on validation data between baseline Transformer model and multi-scale transformer model

	Validation data			
	Baseline transformer		Multi-scale transformer	
	F-score(%)	ER	F-score(%)	ER
without post processing				
NEB=1	25.36	2.37	27.30	2.44
NEB=2	28.77	2.31	28.20	2.33
NEB=3	25.41	2.33	28.08	2.30
Average	**26.51**	**2.34**	**27.79**	**2.35**
Relative improvement	-	-	4.8%	0%
with post processing				
NEB=1	35.44	1.32	37.12	1.26
NEB=2	38.14	1.30	37.91	1.27
NEB=3	35.47	1.33	37.64	1.26
Average	**36.35**	**1.32**	**37.50**	**1.27**
Relative improvement	-	-	3.2%	6.2%

NEB=2일 때는 베이스라인 트랜스포머가 멀티-스케일 트랜스포머보다 F-Score가 더 높지만, NEB=1일 때는 성능이 크게 저하되어 평균 F-score가 더 낮게 나오는 것을 알 수 있었다. 후처리를 사용할 때도 비슷한 결과를 관찰할 수 있었으며 ER에서는 성능 향상이 충분히 크지 않은 것을 확인하였다.

표 2는 평가 데이터를 사용했을 때 SED 결과를 보여주며, 표 1과 대부분 유사한 경향을 관찰할 수 있었다.

표 2. 베이스라인 트랜스포머 모델과 멀티-스케일 트랜스포머 모델 간 평가 데이터에 대한 이벤트 기반 SED 결과 비교

Table 2. Comparison of event-based SED results on evaluation data between baseline transformer model and multi-scale transformer model

	Evaluation data			
	Baseline transformer		Multi-scale transformer	
	F-score(%)	ER	F-score(%)	ER
without post processing				
NEB=1	29.68	1.80	33.13	1.75
NEB=2	34.89	1.69	33.83	1.69
NEB=3	29.29	1.76	35.81	1.61
Average	**31.29**	**1.75**	**34.26**	**1.68**
Relative improvement	-	-	9.5%	4%
with Post Processing				
NEB=1	36.36	1.17	39.32	1.10
NEB=2	41.07	1.11	39.91	1.11
NEB=3	36.52	1.17	40.88	1.08
Average	**37.98**	**1.15**	**40.03**	**1.10**
Relative improvement	-	-	5.4%	4.3%

후처리 과정을 거치지 않은 베이스라인 트랜스포머의 평균 F-score는 31.29%인 반면 멀티-스케일 트랜스포머는 34.26%를 나타내었으며, 상대적인 성능 향상이 9.5%로서 표1의 검증 데이터에 비해 증가하였다. 표 1, 표 2 모두에서 후처리 과정을 적용하지 않았을 때 멀티-스케일 특징의 효과가 더 많이 나타나는 것을 확인할 수 있었다.

Ⅳ. 결 론

본 논문에서는 음향 이벤트 탐지를 위한 딥러닝 멀티-스케일 트랜스포머 모델을 제안하였다.

베이스라인 트랜스포머 모델에 비해 멀티-스케일 특징을 사용한 트랜스포머 모델이 평가 데이터에 대해서 약 9.5%의 상대적인 F-score 향상을 보였다.

또한 멀티-스케일 특징을 사용함으로서 트랜스포머 모델의 인코더 블록의 수가 1에서 3까지 변화함에 따른 성능 변동이 줄어드는 것을 확인할 수 있었다.

실험을 통해 제안된 멀티-스케일 특징이 음향 이벤트 탐지를 위한 트랜스포머 모델의 성능 향상에 기여한다는 것을 확인할 수 있었으며 향후, 많은 양의 레이블이 없는 데이터를 사용할 수 있는 장점으로 현재 SED에서 널리 쓰이는 평균 교사 기반 모델에서도 성능 향상이 나타날 것으로 기대되며 향후 연구과제로 추진할 예정이다. 본 연구의 실험은 DCASE 2019 Challenge의 데이터만을 사용하여 수행되었지만, 관찰된 결과는 음향 이벤트 탐지를 위한 트랜스포머 모델에서 멀티-스케일 특징의 중요성을 나타내기 충분한 것으로 보인다.

References

[1] N. Turpault, R. Serizel, A. Shah, and J. Salamon, "Sound event detection in domestic environemnts with weakly labeled data and soundscape synthesis", Workshop on Detection and Classification of Acoustic Scenes and Events, New York, United States, pp. 253-257, Oct. 2019. https://doi.org/10.33682/006b-jx26.

[2] M. K. Nandwana, A. Ziaei, and J. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, pp. 161-165, Apr. 2015. https://doi.org/10.1109/ICASSP.2015.7177952.

[3] M. Crocco, M. Cristani, A. Trucco, and V.

Murino, "Audio surveillance: a systematic review", ACM Comput. Surv, Vol. 48, No. 52, pp. 1-46, Feb. 2016. https://doi.org/10.1145/2871183.

[4] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks", Adv. Neural Inf. Process, pp. 1097-1105, 2012.

[5] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural Networks", In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 6645-6649, May 2013. https://doi.org/10.1109/ICASSP.2013.6638947.

[6] K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724-1734, Oct. 2014. https://doi.org/10.48550/arXiv.1406.1078.

[7] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection", IEEE/ACM Trans. Audio and Speech Language Process, Vol. 25, No. 6, pp. 1291-1303, May 2017. https://doi.org/10.1109/TASLP. 2017.2690575.

[8] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multilabel deep neural networks", In International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, pp. 1-7, Jul. 2015. https://doi.org/10.1109/IJCNN.2015.7280624.

[9] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks", IEEE/ACM Trans. Audio, Speech and Language Process, Vol. 23, pp. 540-552, 2015. https://doi.org/10.1109/TASLP.2015.2389618.

[10] R. Lu, Z. Duan, and C. Zhang, "Multi-scale recurrent neural networks for sound event detection", In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 131-135, May 2018. https://doi.org/10.1109/ICASSP.2018.8462006.

[11] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pre-defined convolutional neural networks for music auto-tagging", IEEE Signal Process. Letters, Vol. 24, pp. 1208-1212, 2017. https://doi.org/10.1109/LSP.2017.2713830.

[12] X. Li, H. Xianyu, J. Tian, W. Chen, F. Meng, M. Xu, and L. Cai, "A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction", In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Shanghai, China, pp. 544-548, Mar. 2016. https://doi.org/10.1109/ICASSP.2016.7471734.

[13] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117-2125, Jul. 2017.

[14] C. Y. Koh, Y. S. Chen, Y. W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks", ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 376-380, Jun. 2021. https://doi.org/10.1109/ICASSP39728.2021.9414350

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", In 31st Conference on Neural Information Processing Systems, pp. 5998-6008, Dec. 2017.

[16] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution -augmented Transformer for speech recognition",

In Proceedings of the International Conference on Spoken Language Processing (Interspeech), pp. 5036-5040, Oct. 2020.

[17] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Weakly -supervised sound event detection with self-attention", In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 66-70, May 2020. https://doi.org/10.1109/ICASSP40776.2020.9053609.

[18] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer -based sound event detection with semi-supervised learning and data augmentation", Workshop on Detection and Classification of Acoustic Scenes and Events, Tokyo, Japan, Nov. 2020.

[19] J. Gemmeke, D. Ellis, D. Feedman, A. Jasen, W. Lawrence, R. Moore, M. Plakal, and M. Ritter, "Audio set: an ontology and human -labeled dataset for audio events", In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, pp. 776-780, Mar. 2017.

[20] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets", In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR). Suzhou, China, pp. 486–493, Oct. 2017.

[21] G. Dekkers, S. Lauwereins, B. Thoen, M. Adhana, H. Brouckxon, B. Bergh, T. Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network", In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), pp. 32-36, Nov. 2017.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization", In Proceedings of 3rd International Conference for Learning Representations, San Diego, 2015.

저자소개

김 수 종 (Soo-Jong Kim)

2019년 : 계명대학교 전자공학과 (공학사)
2021년 ~ 현재 : 계명대학교 전자전기공학과 석사과정
관심분야 : 인공지능, 오디오 검출

정 용 주 (Yong-Joo Chung)

1988년 : 서울대학교 전자공학과 (공학사)
1990년 : 한국과학기술원 전기및전자공학과(공학석사)
1995년 : 한국과학기술원 전기및전자공학과(공학박사)
1999년 ~ 현재 : 계명대학교 전자공학과 교수
관심분야 : 음성인식, 멀티미디어신호처리