Regular paper
Journal of the KIECS. pp. 389-396, vol. 14, no. 2, Apr. 30. 2019, t. 94, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2019.14.2.389

딥 뉴럴네트워크 기반의 소리 이벤트 검출

정석환* · 정용주**

Sound Event Detection based on Deep Neural Networks

Suk-Hwan Chung* · Yong-Joo Chung**

요 약

본 논문에서는 다양한 구조의 딥 뉴럴 네트워크를 소리 이벤트 검출을 위하여 적용하였으며 공통의 오디오 데이터베이스를 이용하여 그들 간의 성능을 비교하였다. FNN, CNN, RNN 그리고 CRNN이 주어진 오디오데이터베이스 및 딥 뉴럴 네트워크의 구조에 최적화된 하이퍼파라미터 값을 이용하여 구현되었다. 구현된 방식 중에서 CRNN이 모든 테스트 환경에서 가장 좋은 성능을 보였으며 그 다음으로 CNN의 성능이 우수함을 알 수 있었다. RNN은 오디오 신호에서의 시간 상관관계를 잘 추적하는 장점에도 불구하고 CNN 과 CRNN에 비해서 저조한 성능을 보임을 확인할 수 있었다.

ABSTRACT

In this paper, various architectures of deep neural networks were applied for sound event detection and their performances were compared using a common audio database. The FNN, CNN, RNN and CRNN were implemented using hyper-parameters optimized for the database as well as the architecture of each neural network. Among the implemented deep neural networks, CRNN performed best at all testing conditions and CNN followed CRNN in performance. Although RNN has a merit in tracking the time-correlations in audio signals, it showed poor performance compared with CNN and CRNN.

키워드

Deep Neural Networks, Sound Event Detection, Convolutional Recurrent Neural Networks
딥 뉴럴 네트워크, 소리 이벤트 검출, 컨벌루셔널 리커런트 뉴럴 네트워크

Ⅰ. 서 론

소리 이벤트 검출은 패턴 인식 기법을 사용하여 우리 일상에서 발생하는 다양 음향 이벤트들을 찾아내는 기법이다. 이러한 다양한 음향 이벤트들에는 유리창 깨지는 소리, 아기 울음소리, 사람들의 비명소리 및 차의 경적소리들이 포함된다. 소리 이벤트 검출에서는 각 소리들의 종류를 분별하는 것 외에도 발생한 음향 이벤트의 시작 시점과 끝나는 시점도 찾아 주는데, 최근에는 보안, 멀티미디어 데이터로 부터의 정보 추출, 헬스 케어 및 자율주행차 등에 이르기 까지 다양한 분야에서 활용이 가능한 것으로 알려지면서 많은 사람들의 관심을 받게 되었다 [1-5].

지난 수년간 딥 뉴럴 네트워크(DNN: Deep Neural Network)은 영상 분류, 음성인식 그리고 기계 번역 등에 있어서 큰 성공을 거두었으며 [6-8], 최근에 들어서는 딥 뉴럴 네트워크는 이러한 모든 분야에서 최고의 성능을 내는 것으로 알려져 있다. 소리 이벤트

* 계명대학교 전기전자융합시스템공학과(mester88@naver.com)
** 교신저자 : 계명대학교 전자공학과
• 접 수 일 : 2019. 02. 07
• 수정완료일 : 2019. 03. 12
• 게재확정일 : 2019. 04. 15
• Received : Feb. 07, 2019, Revised : Mar. 12, 2019, Accepted : Apr. 15, 2019
• Corresponding Author : Yong-Joo Chung
Dept. Electronic Engineering, Keimyung University,
Email : yjjung@kmu.ac.kr

검출에서도 FNN(: Feedforward Neural Network)은 GMM(: Gaussian Mixture Model)이나 SVM(: Support Vector Machine)에 비해서 우수한 성능을 보임을 알 수 있었으며[9] 기존의 전통적인 방법을 대체한 것으로 보인다.

FNN은 다른 딥 뉴럴 네트워크에 비해서 구조가 간단하여 더 적은 수의 파라미터를 가지고 있으며 계산량이 작다는 장점이 있다. 그러나 FNN의 구조는 입력과 히든층간의 고정된 연결로 인하여 영상 신호에서 발생되는 위치 이동 변이를 보상하기에는 다소 부적합하다. 이와 비슷한 문제가 소리 이벤트 검출에서도 발생하는데, 이는 시간과 주파수 축 상에서 발생하는 소리 신호의 변이를 FNN으로는 충분히 모델링하기가 어렵기 때문이다. FNN을 사용할 경우에 발생하는 또 다른 문제는 입력 소리 신호에 대한 시간 축 상의 콘텍스트(context) 정보가 매우 짧은 시간에 대해서만 고려된다는 것이다. 충분히 긴 시간 동안의 콘텍스트 모델링을 통하여 보다 향상된 성능을 유도할 수 있는 기법이 소리 이벤트 검출에서 요구된다.

FNN에 비하여, CNN(: Convolutional Neural Network)은 시간과 주파수 축 상에서의 소리 신호의 변이를 보다 효과적으로 대처할 수 있으나 CNN도 여전히 소리 신호의 긴 시간동안의 상관관계 정보를 모델링 하는데 있어서는 부족하다. 반면에 RNN(: Recurrent Neural Network)은 음성인식에서 시간 상관관계를 모델링하는데 있어서 매우 성공적이었다. 그러나 RNN은 시간과 주파수 축 상에서는 변이를 효과적으로 모델링 하는데 있어서 CNN에 비하여 불리하며 이런 이유로 CNN에 비해서 소리 이벤트 검출에서 불리하다고 알려져 있다.

CNN과 RNN의 장점을 효과적으로 활용하기 위하여 이들을 결합하고자 하는 많은 연구 노력이 있어 왔다. 최근에는 CNN과 RNN을 하나의 네트워크로 결합한 CRNN(: Convolutional Recurrent Neural Network) 방식이 소리 이벤트 검출뿐만 아니라 음성인식과 음악 분류를 위하여 제안되기도 하였다[10-12].

본 논문에서는 앞에서 언급된 FNN, CNN, RNN 및 CRNN을 이용하여 소리 이벤트 검출을 위한 인식 실험을 실시하고자 하며 이 과정에서 CRNN이 다른 딥 뉴럴 네트워크에 비해서 어떤 장점을 가지고 있는지 심도 있는 조사를 하고자 한다. 또한 CRNN이 최고의 성능을 나타내기 위한 하이퍼파라미터 값과 학습 방식을 도출하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 소리 신호에 대한 특징추출 방법과 본 논문에서 사용된 다양한 딥 뉴럴 네트워크 구조에 대해서 소개한다. 3장에서는 FNN, CNN, RNN 및 CRNN을 이용한 다양한 실험 결과를 비교하며 마지막으로 4장에서 결론을 맺는다.

II. 관련연구

2.1 특징 추출

로그-멜-필터뱅크(log-mel filterbank) 값이 본 논문에서 사용된 딥 뉴럴 네트워크의 특징으로 적용되었다. 먼저 Short-Time Fourier Transform (: STFT)을 44.1 KHz로 샘플링된 오디오 신호의 40ms 구간에 대서 계산하게 된다. STFT 계산 시에는 50%(20ms)의 오버랩을 유지하며, STFT의 결과로부터 40 밴드의 mel-filterbank 값을 구하게 된다. 구해진 멜-필터뱅크 값에 대해서 로그-변환을 적용함으로서 매 20ms 구간마다 40 차원의 로그-멜-필터뱅크 값이 얻어지며 이를 딥 뉴럴 네트워크들의 입력 특징으로 공통적으로 사용한다. 한편 로그-멜-필터뱅크 값들은 직접 사용하는 대신 정규화 과정을 거치게 된다. 먼저 전체 학습 데이터로부터 얻어진 평균값을 로그-멜-필터뱅크 값에서 빼 주고, 역시 전체 학습 데이터로부터 얻어진 표준 편차 값으로 나누어 줌으로서 딥 뉴럴 네트워크의 입력으로 사용되는 특징에 대한 정규화가 이루어지게 된다.

2.2 FNN

40차의 로그-멜-필터뱅크 특징은 연속적인 5개의 프레임들이 합쳐져서 200차원의 특징벡터를 형성하며 이들은 FNN의 입력으로 사용된다. 또한 2개의 은닉층은 각각 ReLu 활성함수를 가지는 1600개의 유닛들로 구성된다. 출력 층은 분류하고자 하는 소리 이벤트 클래스의 수만큼의 유닛을 가지며 각 유닛은 sigmoid 활성함수를 이용한다. 여기서 sigmoid 활성함수의 출력값은 각 클래스에 대한 사후확률로 간주되며 0.5의

기준치와 비교하여 이진화 된 후 ground truth table과 비교하여 FNN의 정확도(accuracy) 산출을 위하여 사용된다.

2.3 CNN

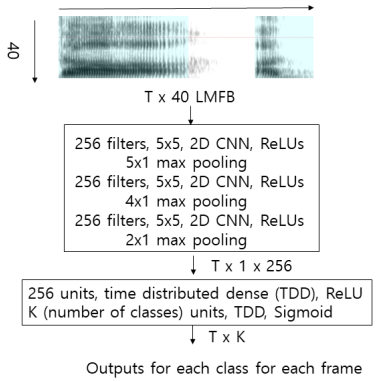


그림 1. CNN 구조

Fig. 1 CNN Architecture

그림 1에 나타난 바와 같이 CNN의 구조는 컨벌루션(convolution) 층, max pooling 층 그리고 단일 층의 FNN 및 출력 층으로 이루어져 있다. 먼저, T 프레임의 40차의 ($T \times 40$) 로그-멜-필터뱅크 값이 CNN의 컨벌루션 층의 입력으로 사용된다. 컨벌루션 층은 256개의 feature map으로 이루어져 있으며 5x5의 2차원 컨벌루션 필터를 사용하여 입력 층으로부터 특징을 받아들인다. 또한, 컨벌루션 층의 각 유닛은 ReLu 활성함수를 가진다. 컨벌루션 층의 출력은 겹쳐짐이 없는 max pooling 층을 거침으로서 데이터의 차원 감소가 이루어진다. 하지만 데이터의 시간 정보를 보존하기 위하여, max pooling은 주파수 영역에서만 이루어지게 된다. 이것은 영상 분류에 사용되는 CNN에서 차원 감소가 2차원 모두에서 이루어지는 것과는 대조적인데, 이는 소리 이벤트 검출을 위해서는 이벤트의 시작점과 끝점을 찾기 위해서 오디오 신호의 시간 해상도 정보가 유지되어야하기 때문이다. 여기서는 3개의 컨벌루션 층이 사용되었으며 컨벌루션 층과 max pooling 층을 거친 후의 특징의 차원은 $T \times 1 \times 256$ 이 된다. 앞서 언급된 바와 같이 시간 영역의 차원은 T로 유지된 반면 주파수 차원은 1로 줄어드는 것을 확인 할 수 있다. Max pooling을 거친 후 생성된 특징은 256개의 유닛을 가진 단일 FNN을 통과하고 마지막으로 sigmoid 활성 함수를 가진 출력 층에 입력됨으로써 CNN의 전체적인 네트워크 구성이 설계된다. 출력 층의 sigmoid 활성함수의 결과 값은 0.5의 임계치를 이용하여 이진화 되며 이를 바탕으로 각 시간 프레임에서의 특정 클래스의 소리 이벤트의 활성화 여부가 결정된다.

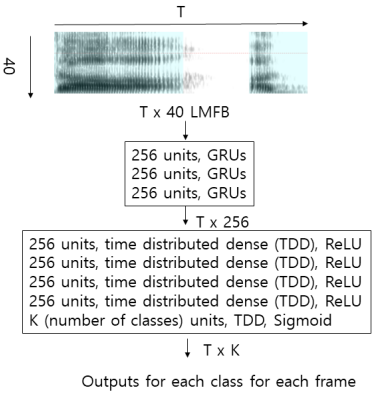


그림 2. RNN 구조

Fig. 2 RNN Architecture

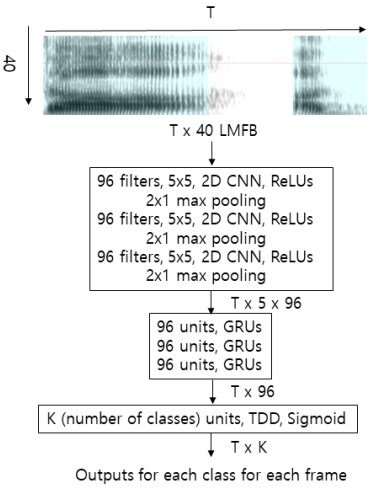


그림 3. CRNN 구조

Fig. 3 CRNN Architecture

2.4 RNN

RNN의 입력으로는 CNN과 동일한 T 프레임의 40

차의 ($T \times 40$) 로그-멜-필터뱅크 값이 사용되며, 그 전체적인 구조는 256개의 유닛을 가진 3개의 GRU (: Gated Recurrent Unit)층과 4개 층의 FNN 그리고 마지막에 sigmoid 활성함수를 가진 출력 층으로 구성되어 있으며 보다 자세한 내용은 그림 2에 나타나 있다. 4개 층의 FNN을 사용함으로서 CNN 과 RNN이 비슷한 수준의 깊이를 가지게 하며 두 모델들 간의 성능 비교가 정당화 될 수 있다고 생각된다.

2.5 CRNN

본 논문에서 사용된 CRNN의 구조가 그림 3에 나타나 있다. 3개의 컨벌루션 층과 3개의 GRU 층 그리고 마지막으로 출력 층이 연속적으로 연결된 형태를 띠고 있다. 컨벌루션 층은 시간과 주파수 영역에서 변이에 강인한 특징 추출의 역할을 하며, GRU 층은 시간 영역에서의 상관관계 정보를 제공하는 역할을 하는데 이는 소리 이벤트 검출에서 매우 중요하다. 마지막의 출력 층은 주어진 시간 프레임에서의 소리 이벤트 클래스에 대한 사후 확률을 제공하게 되며, 이진화를 통하여 각 클래스에 대한 존재 유무를 판단하는 근거로 사용된다.

III. 실험 방법 및 결과

3.1 실험 방법

본 논문에서는 이 분야의 연구에서 잘 알려진 오디오 데이터베이스인 TUT Sound Events Synthetic 2016 (TUT-SED Synthetic)을 사용하였다[13]. 이 데이터에는 인공적으로 만들어진 오디오 데이터가 포함되어 있는데, 이는 자연적인 녹음을 통하여 얻을 수 있는 데이터의 양의 부족을 보완하기 위함이다. 또한 인위적인 데이터를 사용함으로서 실제 환경에서 만들어진 오디오에 대한 레이블의 산정 시 주관적인 요소가 많이 들어가는 단점도 보완해 줄 수 있다.

TUT-SED Synthetic 데이터베이스는 16개의 서로 다른 클래스를 가진 오디오 샘플 들을 인위적으로 섞어서 만들어진다. 전체 파일의 개수는 100개이며 이들은 994개의 개별적인 오디오 샘플들을 서로 섞어서 만들어진다. 100개의 파일은 학습과 테스트 그리고 검증 데이터로 나누어지며 그 비율은 60:20:20이다. 전체 파일의 길이는 556분으로 비교적 긴 편이다.

3.2 실험 결과

제안된 딥 뉴럴 네트워크에서 최적의 성능을 얻기 위하여 배치정규화(batch normalization)가 모든 컨벌루션 층에 사용되었으며 드랍아웃(dropout)은 모든 컨벌루션 층과 GRU 층에 사용되었으며, 이때 드랍아웃 비율(rate)은 0.25로 하였다. 딥 뉴럴 네트워크의 학습은 바이너리(binary) 크로스엔트로피 (cross-entropy) 비용함수를 이용하여 Adam 최적화를 통하여 이루어졌다. 오버피팅을 방지하기 위하여 조기종료(early stopping)가 학습과정에서 적용되었으며 비용함수 값이 100 epoch 동안 감소하지 않으면 학습을 종료하게 된다.

표 1. Learning rate 변화에 따른 CRNN 성능
Table 1. Performances of CRNN as learning rate changes

learning rate	validation data		testing data		epoch
	segment-based (F-score/ER)	event-based (F-score/ER)	segment-based (F-score/ER)	event-based (F-score/ER)	
10^{-3}	61.6% /0.52	37.6%/ 0.96	60.6% /0.53	37.0%/ 0.97	16
10^{-4}	**68.7% /0.45**	**43.4% /0.88**	**64.2% /0.50**	**40.5% /0.96**	33
10^{-5}	66.4% /0.49	39.1% /0.96	63.7% /0.52	36.4%/ 1.04	157
10^{-6}	44.1% /0.69	9.8% /1.24	43.3% /0.71	10.8%/ 1.27	191

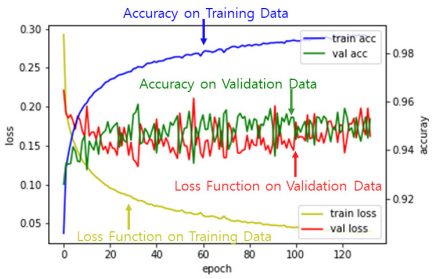
딥 뉴럴 네트워크의 성능은 learning rate에 따라서 다르게 되는데 최적의 learning rate를 찾기 위하여 우리는 검증데이터에 대한 성능 평가를 실시하였다. 표 1에는 learning rate가 달라짐에 따라서 CRNN의 성능이 어떻게 변화하는지를 나타내었다.

보다 신뢰성 있는 실험결과를 도출하기 위하여 평가 방식으로는 segment 기반의 방식과 event 기반의 방식을 모두 사용하였고 평가 척도로는 소리 이벤트 검출에서 대표적으로 많이 사용되는 F-score와 Error Rate를 모두 사용하였다[5].

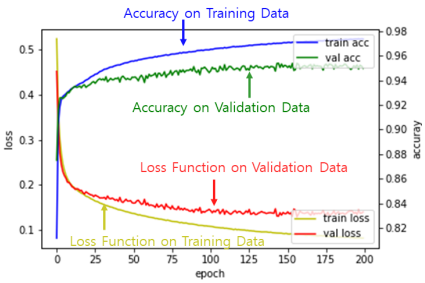
표 1의 결과로부터, learning rate 값이 10^{-4}일 경우에 모든 조건에서 가장 좋은 성능을 보임을 알 수 있다. 또한 검증데이터에서 가장 좋은 성능을 보이는 learning rate값이 인식데이터에서도 가장 좋은 성능을 보임을 알 수 있는데, 이는 검증데이터에 대한 인식 성능을 근거로 하여 learning rate를 결정하는 것이 합리적임을 말해 준다. 비록 본 논문에서는 제시하지 않았지만, CRNN에서와 비슷한 결과들이 FNN, CNN, RNN 등에서도 얻어짐을 추가적인 실험을 통해서 확인할 수 있었다. 표 1에는 최고의 성능을 나타내는 epoch의 횟수가 표시되어 있는데, learning rate가 감소함에 epoch의 수가 증가함을 알 수 있다. 이는 learning rate 값이 작아짐에 따라서 딥 뉴럴 네트워크들이 천천히 수렴함에 따른 것으로 생각된다. 예를 들어, learning rate 가 10^{-4} 일 때, epoch 수는 33 이지만 learning rate 가 10^{-7} 일 때, epoch 수는 191 이 된다. 수렴이 너무 느려지는 경우는 성능의 저하도 야기 시키는데 이는 딥 뉴럴 네트워크의 언더피팅(underfitting)에 의한 것이라 판단된다.

Learning rate에 따른 성능 변화에 대한 보다 깊이 있는 관찰을 위하여 그림4에는 CRNN의 출력 값에 대한 비용함수(loss function) 값과 정확도(accuracy)의 변화를 epoch 횟수에 따라서 그래프로 나타내었다.

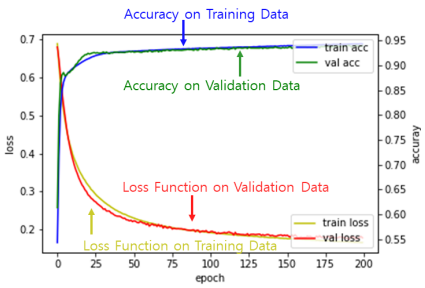
그림 4(a)에서 보다시피, learning rate가 10^{-4} 일 때, 검증데이터에 대한 비용함수의 값은 33 번째 epoch에서 최소값을 갖는 것을 알 수 있으며, 그 이후 비용함수의 값은 다소 출렁이지만 최소값 아래로는 내려오지 않는다. 반면, 학습데이터의 경우에는 학습과정이 끝날 때 (200 epoch) 까지 비용함수의 값은 계속 감소하는 것을 알 수 있다. 그러나 학습과정에서 오버 피팅(overfitting)을 방지하는 것이 매우 중요하므로, 본 논문에서 구현된 조기종료 알고리듬에 의해서 학습은 epoch 33에서 마치게 된다. 한편, 그림 4(b)에서처럼, learning rate가 10^{-5} 인 경우에는, 상당히 다른 현상이 나타남을 그래프 상에서 확인 할 수 있다. 이 경우, 검증 데이터에 대한 비용함수 값은 더 오랜 기간 동안 계속해서 감소하는 것을 볼 수 있는데 epoch 157에서 최소값을 가진다. 이와 같이 더 길어진 epoch 횟수는 언더피팅(underfitting)을 야기 시킴으로써 성능의 저하를 불러 오는 것으로 생각된다. 이러한 현상은 learning rate를 더 증가시킴으로서 더욱 더 명백해지는 것을 알 수 있다. 그림4(d)에서는 learning rate가 10^{-7}으로 매우 작게 설정되었는데, 이로 인하여 비용 함수 값이 마지막 epoch가 될 때까지도 수렴하지 않는 것을 알 수 있으며, 이러한 현상은 정확도(accuracy)에서도 그대로 나타난다.



(a) Learning rate= 10^{-4}



(b) Learning rate= 10^{-5}



(c) Learning rate= 10^{-6}

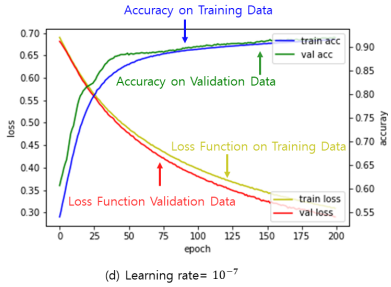


그림 4. Learning rate 변화에 따른 비용함수와 정확도의 epoch에 따른 수렴 특성.
Fig. 4 Convergence Characteristics of Loss Function and Accuracy as Learning rate changes.

표 2에는 FNN, CNN, RNN 과 CRNN에 대한 성능비교가 나타나 있으며 앞의 실험 결과를 반영하여 모든 딥 뉴럴 네트워크의 learning rate는 10^{-4}으로 설정하였다. 표 2의 결과로부터, 모든 테스트 조건에서 CRNN이 가장 우수한 성능을 보임을 알 수 있었다. CNN은 영상인식 분야에서는 꽤 만족스러운 결과를 보이지만 오디오 신호의 긴 시간-상관관계 정보를 모델링 하는데 부족함을 보이기 때문에 소리 이벤트 검출에서는 CRNN에 비해서 열등한 성능을 나타내는 것으로 판단된다. RNN은 시간-상관관계 정보를 나타내는데 있어서 CNN에 비해서 유리하지만 성능 면에서는 오히려 뒤처지는 것으로 나타났다. 이는 CNN이 자생적으로 만들어 내는 시간-주파수 변이에 강인한 특징이 소리 이벤트 검출에 있어서 매우 중요함을 나타내는 것이라 생각된다.

표 2. FNN, CNN, RNN, CRNN 간의 성능비교
Table 2. Performance Comparison Between FNN, CNN, RNN and CRNN.

	Segment-based		Event-based	
	F-score	ER	F-score	ER
FNN	54.5 %	0.8	21.4 %	3.51
CNN	60.3 %	0.66	31.3 %	1.87
RNN	47.2 %	0.66	28.9 %	1.32
CRNN	64.2 %	0.5	40.5 %	0.96

Ⅳ. 결론

딥 뉴럴 네트워크는 패턴 인식의 다양한 분야에서 사용되고 있으며, 기존의 패턴 인식 방식에 비해서 우수한 성능을 보여 주었다. 본 논문에서는 최신의 딥 뉴럴네트워크 방식을 소리 이벤트 검출에 적용하고 최적의 인식 성능을 얻기 위한 방법에 대해서 논의하였다.

본 논문에서는 FNN, CNN, RNN 그리고 CRNN을 소리 이벤트 검출에 적용하고 성능을 비교하였다. 딥 뉴럴 네트워크의 성능은 learning rate에 상당히 좌우되는 것을 알 수 있었으며, 최적의 learning rate 값은 딥 뉴럴 네트워크의 종류에 따라서 달라지지 않음을 확인할 수 있었다. 또한 learning rate값을 검증 데이터의 성능을 기준으로 정하면 테스트 데이터에 대해서도 최적의 성능을 보임을 확인 할 수 있었다. 너무 작은 값의 learning rate는 딥 뉴럴 네트워크를 학습 데이터에 대해서 언더피팅 하도록 하며 반면에 너무 큰 값은 오버피팅 하도록 하여 성능을 저하를 발생시키는 것을 확인 할 수 있었다.

본 논문에서 실험한 다양한 딥 뉴럴 네트워크 중에서 CRNN이 음향 이벤트 검출에서 가장 좋은 성능을 보임을 알 수 있었으며, CNN이 그 뒤를 이었다. CNN의 성능이 오디오 신호에 대한 시간-관계 정보를 잘 표현하는 RNN에 비해서 오히려 우수한 성능을 보였는데, 이를 통해서 CNN에서 추출되는 시간-주파수 변이에 강인한 특징이 매우 중요함을 확인 할 수 있었다.

향후 음향 이벤트 검출의 보다 나은 성능을 위해서는 CRNN 구조의 다양한 변형에 대한 개발이 필요하며, 배치 정규화나 드랍아웃에 대한 최적의 적용 방법 등에 대한 연구가 추가적으로 필요할 것이라 생각된다. 또한 CRNN의 입력으로 사용되는 오디오 세그먼트의 길이의 최적 값에 대한 조사도 필요하리라 판단된다.

감사의 글

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임.(NRF-2018R1A2B6009328)

References

[1] M. 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, April, 2015.

[2] M. Crocco, M. Christani, A. Trucco, and V. Murino, "Audio Surveillance: A Systematic Review," *ACM Computing Surveys,* vol. 48. no. 4, 2016, pp. 52:1-52:46.

[3] Y. Lee and P. Moon, "A Comparison and Analysis of Deep Learning Framework," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 12, no. 1, 2017, pp. 115-122.

[4] Y. Wang, L. Neves, and F. Metze, "Audio-based Multimedia Event Detection Using Deep Recurrent Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 2742-2746.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences,* vol. 6, no. 6, 2016, pp. 321-337.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Communications of the ACM,* vol. 60, no. 6, 2017, pp. 84-90.

[7] A. Graves, A. Mohamed, and G. E. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *Proceedings of the IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6645-6649.

[8] S. Bang, "Implementation of Image based Fire Detection System Using Convolution Neural Network," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 12, no. 2, 2017, pp. 331-336.

[9] S. Chung and Y. Chung, "Comparison of Audio Event Detection Performance using DNN," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 13, no. 3, 2018, pp. 571-577.

[10] 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 and Language Process*, vol. 26, no. 6, 2017, pp. 1291-1303.

[11] T. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-term Memory, Fully Connected Deep Neural Networks," *Proceedings of the 2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Brisbane, Australia, 2015, pp. 4580-4584.

[12] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional Recurrent Neural Networks for Music Classification," *Proceedings of the 2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* New Orleans, USA, 2017, pp. 2392-2396.

[13] TUT-SED Synthetic Database 2016, Availab:http://www.cs.tut.fi/sgn/arg/taslp2017-crnn-sed/tut-sed-synthetic-2016

저자 소개

정석환(Suk-Hwan Chung)

2017년 계명대학교 전자공학과 졸업(공학사)
2017년~현재 계명대학교 일반대학원 전기전자융합시스템공학과 석사과정
※ 관심분야 : 인공지능, 오디오 검출

정용주(Yong-Joo Chung)

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