Journal of KIIT. Vol. 17, No. 5, pp. 83-90, May 31, 2019. pISSN 1598-8619, eISSN 2093-7571 83
http://dx.doi.org/10.14801/jkiit.2019.17.5.83

CRNN 기반의 소리 이벤트 검출 성능 분석

정석환*, 정용주**

Performance Analysis of Sound Event Detection Based on CRNN

Suk-Hwan Chung*, Yong-Joo Chung**

이 논문은 한국연구재단의 기초연구사업 지원을 받아 수행된 것임.(NRF-2018R1A2B6009328)

요 약

본 논문에서는 소리 이벤트 검출을 위한 CRNN(Convolutional Recurrent Neural Network)에 대한 심층적 고찰을 진행하였다. 이를 위해서 학습률(learning rate)과 입력 세그먼트의 길이에 대한 최적의 값을 성능 실험을 통하여 알아보았다. 또한, 매 학습 반복(iteration) 마다 배치(batch) 데이터의 변화가 성능에 어떤 영향을 미치는지를 고찰하였으며, 배치정규화(batch normalization)과 드랍아웃(dropout)의 유무가 성능에 미치는 영향에 대해서도 관찰하였다. 마지막으로, 학습 과정의 수렴을 판단하는 최적화 기준에 따른 성능 변화도 알아보았다. 앞에서 언급된 이러한 과정을 통해서 소리 이벤트 검출을 위한 CRNN에 대한 최적의 조건을 도출하고자 하였다. TUT Sound Events Synthetic 2016 데이터베이스를 이용한 실험 결과 학습률은 10^{-4} 일 때 가장 좋은 성능을 보였고 드랍아웃 보다는 배치정규화 성능이 훨씬 큰 영향을 미치는 것을 알 수 있었으며 매 반복 마다 배치 데이터의 시작점을 다르게 함으로써 많은 성능 향상을 이루었다.

Abstract

In this paper, we performed an intensive investigation on the CRNN for audio event detection. For this purpose, we tried to find out the optimal learning rate and input segment length for the training via performance experiments. In addition, we investigated the effects of variation in each iteration of the batch data on the performance and the effects of batch normalization and dropout were also observed. Finally, we figured out the performance variation depending on the optimization criterion for the convergence of training process. Through the aforementioned process, we tried to find out the optimal conditions for the CRNN. From the experimental results using the TUT Sound Events Synthetic 2016 database, we could obtain the best peformance with the learning rate 10^{-4} and batch normalization was far more effective on performances and great performance improvement was obtained by varying the starting point of the batch data.

Keywords
sound event detection, CRNN, optimal learning condition, deep neural networks

* 계명대학교 전자공학과
- ORCID: https://orcid.org/0000-0002-1898-4947
** 계명대학교 전자공학과(교신저자)
- ORCID: https://orcid.org/0000-0002-0060-1178

· Received: Feb. 26, 2019 Revised: Mar. 25, 2019, Accepted: Mar. 28, 2019
· 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

최근에 많은 연구가 진행되고 있는 소리 이벤트 검출에서는 우리가 일상생활에서 자주 접하는 다양한 소리들, 예를 들면, 창문 깨지는 소리, 아기들 울음소리, 비명소리, 자동차의 경적소리 등의 유무를 판단하게 된다. 여기서 한발 더 나아가서, 소리 이벤트 검출에서는 발생한 소리 이벤트의 시작시점과 끝나는 시점도 찾는 것을 시도한다. 소리 이벤트 검출은 감시(Surveillance)나 도심에서 발생하는 소리의 분석, 멀티미디어 컨텐츠 내의 정보 탐색, 숲속의 새 소리 찾기, 자율주행 등의 다양한 응용 분야를 가지고 있으며 최근에 들어서 많은 연구자들의 관심이 되고 있다[1]-[5].

딥 뉴럴네트워크(Deep Neural Networks)는 최근에 들어서 영상 분류나 음성인식, 기계번역 등의 여러 분야에서 기존의 머신러닝 기법들에 비해서 우수한 성능을 보이며 많은 성공을 거두고 있다[6]-[8]. FNN(Feedforward Neural Network)는 소리 이벤트 검출에서 과거에 많이 사용되었던 GMM(Gaussian Mixture Model) 이나 SVM(Support Vector Machine)에 비해서 우수한 성능을 보여 줌으로써 최근에는 전통적인 기법들을 대체할 수 있게 되었다[9]. FNN은 딥 뉴럴네트워크 중에서도 비교적 간단한 구조를 가지며 이로 인하여 계산량이나 메모리 소요량이 적게 소모되는 장점이 있다.

FNN는 입력신호와 은닉층 간의 고정된 연결구조로 인하여 영상 분류에서처럼 2차원 공간에서 발생하는 신호 변이를 극복해야 될 필요성이 있는 경우, 다소 부적합한 것으로 알려져 있다. 영상 신호와 비슷하게 오디오 신호의 경우에도 시간-주파수 공간에서 신호 변이가 빈번히 발생하므로 FNN으로 이를 효과적으로 모델링하기에는 다소 부족한 것으로 판단된다. 또한, 오디오 신호의 경우에는 시간프레임 간의 상관관계를 효과적으로 모델링 하는 것이 중요한데 FNN의 경우에는 단순히 입력신호에서 몇 개의 프레임을 연결함으로써 시간프레임 상관관계를 나타내므로 충분히 긴 시간 동안의 입력 샘플들 간의 상관관계를 모델링하기에는 부족한 것으로 생각된다.

CNN(Convolutional Neural Networks)은 사람의 시각 피질 (visual cortex)에서 영상의 일부분에 대해서만 반응하는 활동을 모사한 구조로 다양한 패턴인식 문제에서 FNN보다 우수한 성능을 보여 주었다. 특히, CNN은 그 구조의 특성상 영상신호의 2차원 공간 변이를 효과적으로 모델링 하는 것으로 알려져 있으며, 이러한 성질이 오디오 신호의 분류에도 효과적임이 알려져 있다. 그러나 CNN의 경우에는 오디오 신호에서 발생하는 긴 시간동안의 샘플들 간의 상관성을 모델링하는데에 있어서는 충분치 못한 것으로 판단된다.

RNN(Recurrent Neural Networks)은 음성인식에서 성공적으로 사용되었으며 음성 및 오디오 신호의 시간 상관관계를 모델링 하는데 있어서 우수하다고 알려졌다 . 그러나 RNN은 오디오 신호의 시간-주파수의 2차원 공간에서의 변이를 모델링 하는데 있어서 다소 부족하기 때문에 단독으로 사용될 경우, 소리 이벤트 검출에서는 CNN보다 열등한 성능을 보여주는 것으로 알려져 있다.

최근에 들어서는 CNN과 RNN 각각의 장점을 모두 이용하기 위하여 이들을 결합하고자 하는 시도가 다수 있었다. 그중에서도 CRNN(Convolutional Recurrent Neural Network)은 CNN과 RNN을 직렬 연결하여 하나의 네트워크를 구성하도록 설계되었으며 소리 이벤트 검출이나 음성인식, 음악 분류 등에서 사용된다[10]-[12]. CRNN은 CNN, RNN 그리고 FNN이 함께 연결되어 만들어진 구조이므로 그들이 단독으로 사용될 경우에 비해서 구조가 복잡하며 상호 연동된 결과가 정확히 어떤 작용을 할지 예측하기가 어려운 점이 있다. 또한, 비교적 최근에 이르러서 CRNN을 이용한 소리 이벤트 검출 연구가 시작되었으므로, 학습 과정에서 CRNN에서 사용되는 다양한 파라미터들을 최적화하고자 하는 연구들이 다소 부족하였다. 따라서 본 연구에서는 소리 이벤트 검출에서 사용되는 CRNN에 적용되는 조건들을 다양하게 실험함으로써 기존의 딥 뉴럴네트워크에서 일반적으로 사용되었던 하이퍼파라미터들이 CRNN에서도 충분히 적절한지 파악하고 이를 통해서 CRNN에 대한 최적의 학습 조건을 설정함으로써 소리 이벤트 검출 성능의 향상을 도모하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 소리 이벤트 검출을 위한 특징 추출 방법과 본 논문에서 사용된 CRNN 인식기의 구조에 대해서 소개하며 3장에서는 본 연구에서 수행한 다양한 인식 실험결과를 제시하고 4장에서 결론을 맺는다.

II. 특징 추출과 CRNN 구조

2.1 특징 추출

본 연구에서는 로그-멜-필터뱅크(Log-mel Filterbank) 값을 CRNN의 입력 특징으로 사용하였으며, 전체적인 추출과정은 그림 1에 나타나 있다.

먼저, 44.1KHz로 샘플링된 오디오 신호의 매 40ms 구간에 대해서 20ms의 중첩을 허용하여 STFT (Short-Time Fourier Transform)을 구한다. STFT 값으로부터 40 band의 멜-필터뱅크 값을 전체 0에서 22050Hz의 주파수 구간에서 구한 후, 이를 로그변환 함으로써 40차원의 로그-멜-필터뱅크 값이 20ms 간격의 프레임마다 얻어지게 된다. 로그-멜-필터뱅크 값은 학습 데이터 전체의 평균값으로 빼주고 또한 표준편차 값으로 나누어주는 과정을 통해서 정규화한 후 사용하게 된다.

2.2 CRNN의 구조

본 논문에서 사용된 CRNN의 구조가 그림 2에 나타나 있다. CRNN은 CNN와 그 뒤를 따라오는 RNN 그리고 출력을 나타내는 FNN이 서로 연결된 형태로 이루어진다. CNN은 시간-주파수 공간 상의 변이에 강인한 특징을 추출하는 역할을 하며, RNN은 시간영역에서의 오디오 신호의 상관관계 정보를 제공하게 된다. 마지막으로 FNN은 특정 시간 프레임에서 각 소리 이벤트들에 대한 사후 추정 확률값 (Posterior Probabilities)을 출력한다.

CNN의 경우에는 오디오 신호로부터 추출된 40 차원의 로그-멜-필터뱅크 값을 입력으로 받아들이는데, 이때 한 번에 입력되는 시간 프레임의 길이(T)는 최적의 인식 성능을 나타내기 위하여 조절이 가능하므로 CNN의 입력 특징의 차원은 $T \times 40$ 이 된다.

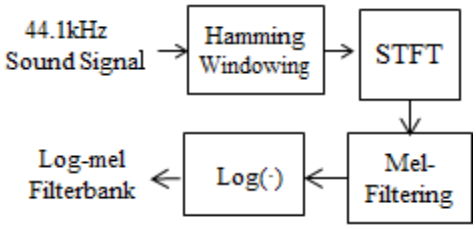

그림 1. 로그-멜-필터뱅크 추출 과정
Fig. 1. Extraction process of log-mel filterbank

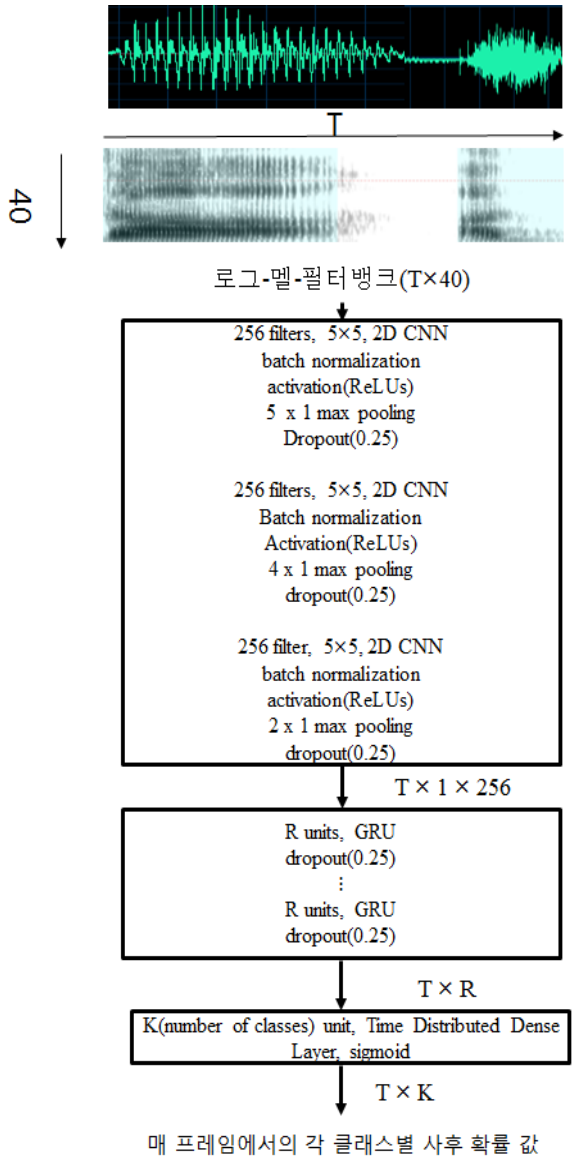

그림 2 CRNN의 구조
Fig. 2 Architecture of CRNN

CNN은 3개의 컨벌루션 층으로 이루어져 있으며, 각 층은 독립적인 5×5 필터를 사용하는 256개의 특징 맵(Feature Map)으로 구성된다. 필터의 출력에 대해서는 배치정규화를 적용한 후 ReLU 활성화 함수

를 거치게 된다. 시간축 상에서의 오디오 신호 정보를 유지하기 위해서 CNN에서의 최대 풀링(Max Pooling)은 주파수 축 상에 서만 이루어진다. 각 컨벌루션 층의 마지막 부분에서는 0.25의 비율로 드랍아웃이 실행된다.

CNN의 출력은 GRU(Gated Recurrent Units)를 기본 구성 단위(Unit)로 채택한 RNN의 입력으로 사용된다. RNN은 256개의 GRU 유닛을 가지며 1개 이상의 층으로 구성된다. 최종적으로 출력층은 sigmoid 함수를 활성함수로 가지는 K(오디오 신호 클래스의 수)개의 단위를 가지며 매 프레임 마다 각 클래스의 사후 추정 확률값을 생성하며 이를 바탕으로 CRNN의 소리 이벤트 검출 성능이 결정된다.

III. 실험 결과

3.1 데이터베이스 및 평가척도

본 논문에서는 CRNN의 성능 실험을 위하여 이 분야에서 많이 활용되는 대표적인 오디오데이터베이스인 TUT Sound Events Syhtnetic 2016 (TUT-SED Synthetic)을 이용하였다[13]. 이 데이터에는 인공적으로 만들어진 오디오 데이터가 포함되어 있는데, 이는 실제 환경에서 녹음된 오디오만을 이용할 경우 데이터가 충분히 확보되기 어렵기 때문이다. 또한 인위적으로 오디오 데이터를 만들어 줌으로써 레이블의 산정시 주관적인 요소가 많이 들어가는 단점을 보완해 줄 수도 있다는 장점이 있다.

TUT-SED Synthetic은 16개의 서로 다른 종류(클래스)로부터의 소리 이벤트들을 섞어서 만들어진다. 데이터의 전체 길이는 566분이며 이들을 60%, 20%, 20%의 비율로 나누어 각각 학습데이터, 검증데이터, 인식데이터로 활용하였다. 소리 이벤트의 종류는 16개이며 표 1에 그들의 구체적인 종류와 전체 길이가 표시되어 있다.

표 1. TUT-SED Synthetic 2016 데이터베이스의 소리 종류와 전체 시간
Table 1. Sound classes and their total duration in the TUT-SED synthetic 2016 databases

classes	duration(s)	classes	duration(s)
glass	621	motor cycle	3691
gun	534	foot steps	1173
cat	941	claps	3278
dog	716	bus	3464
thunder	3007	mixer	4020
bird	2298	crowd	4825
horse	1614	alarm	4405
crying	2007	rain	3975

CRNN을 이용한 소리 이벤트 검출의 평가척도로는 F-score와 ER(Error Rate)를 동시에 사용하였으며, 평가 방식으로는 테스트 오디오의 1초 단위마다 평가척도를 계산하는 세그먼트(Segment)기반 방식과 실제 이벤트가 발생할 때 마다 평가척도를 계산하는 이벤트(Event) 기반 방식을 모두 사용하였다[14].

3.2 인식 결과

본 연구에서는 CRNN의 학습과정에서 다양한 환경을 적용하여 최적의 인식성능을 나타내는 하이퍼-파라미터 값을 찾고자 하였다. 먼저 학습률을 변화시키면서 소리 이벤트 검출의 성능 변화를 관찰하였으며 그 결과는 표 2에 나타나 있다.

표 2. 학습률 변화에 따른 CRNN 성능
Table 2. Performances of CRNN as learning rate changes

learning rate	validation data		testing data		epoch
	segment method (F-score/ER)	event method (F-score/ER)	segment-method (F-score/ER)	event method (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

성능 실험을 위해서 bach normalization과 드랍아웃을 모두 적용하였다. 비용함수로는 binary cross entropy를 사용하였으며 이는 학습의 수렴여부를 판단하는 기준이 된다. 한편, 뉴럴네트워크의 최적화는 Adam optimizer를 사용하였다.

표 2에서 보다시피, CRNN에서의 최고의 성능은 학습률이 10^{-4}인 경우에 나타남을 알 수 있다. 일반적으로 딥 뉴럴네트워에서는 최적의 학습률이 네트워크의 형태나 학습데이터의 양에 따라서 다소 달라지기 때문에 본 연구에서는 검증데이터에 대한 성능을 기준으로 학습률을 결정하였다. 표 2에 나타나 있듯이 검증데이터에서 최고의 성능을 보이면 인식데이터에서도 가장 좋은 성능을 보임을 알 수 있으므로, TUT-SED Synthetic을 이용한 CRNN의 학습에서는 검증데이터를 이용한 학습률 선택이 적절함을 알 수 있다.

표 2에서는 학습률이 감소함에 따라서 최고의 성능을 나타나는 반복 횟수가 증가함을 할 수 있다. 예를 들어, 학습률이 10^{-4}인 경우에는 33회의 반복 횟수를 나타내지만 10^{-6}인 경우에는 191회의 반복 횟수를 가지는데, 이는 CRNN의 weight들이 느리게 수렴하기 때문이며 CRNN의 underfitting을 야기하고 성능저하를 낳는다. 반면에 학습률이 10^{-3}으로 커지게 되면 반복횟수가 16회로 급격히 줄어들며 이는 overfitting을 야기시킴으로써 또한 성능의 저하를 야기한다.

본 연구에서는 훈련에 사용되는 오디오 세그먼트들을 매 epoch 마다 변화를 준 경우에 인식율의 향상이 있는지 조사해 보았다. 이를 위해서는 매 epoch마다 오디오 세그먼트들의 시작 프레임을 순차적으로 이동시킴으로써 CRNN이 받아들이는 세그먼트들이 매 epoch 마다 달라지게 하는 방식을 취하였다(Overlap방식).

표 3에는 위와 같은 overlap 방식을 적용한 경우(Overlap)와 그렇지 않은 경우(Non-overlap)에 대해서 소리 이벤트 검출결과를 나타내었다. 평가방식으로는 세그먼트 기반 방식과 이벤트 기반 방식을 모두 사용하였으며 평가결과는 F-score와 ER로 나타내었다. 학습의 수렴 기준은 검 증데이터에 대한 ER 값을 적용하였으며 조기 종료 과정을 통하여 수렴 기준 값이 100회(Iteration) 동안 개선되지 않으면 overfitting을 방지하기 위해서 학습을 종료하며 최대 학습 회수는 200으로 설정하였다. 세그먼트 길이 T는 1024 프레임(20.48초)으로 정하였다.

또한 본 연구에서는 배치정규화와 드랍아웃의 영향에 대해서도 조사하였다. 배치정규화는 학습의 역전파(Backpropagation) 과정에서 발생하는 vanishing gradient 및 exploding gradient 문제를 해결하기 위하여 사용되는 기법이며, 본 연구에서는 그림 2에서 보인바와 같이 컨벌루션 층에서 ReLU 활성함수 전단계에서 적용된다. 드랍아웃은 뉴럴네트워크의 정형화(Regularization) 기법중의 하나로, 출력층을 제외한 각 층의 뉴런을 사전에 정의한 확률(Dropout Rate)로 학습 과정에서 배제시키는 기법이다. 본 연구에서는 컨벌루션 층과 RNN층 모두에 드랍아웃을 적용하였다. 표 2에는 배치정규화와 드랍아웃을 적용했을 경우와 하지 않았을 경우 각각에 대한 인식성능을 표시하였다.

표 3 학습조건의 변화에 따른 CRNN 성능
Table 3. Performances of CRNN as training conditions changes

batch normalization	dropout	segment-based method		event-based method		
		Overlap (F-score/ER)	Non-overlap (F-score/ER)	Overlap (F-score/ER)	Non-overlap (F-score/ER)	average
yes	no	66.10%/0.48	65.28%/0.54	43.80%/1.01	42.99%/1.12	**54.54%/0.79**
yes	yes	67.24%/0.49	66.62%/0.49	45.93%/0.94	46.88%/0.92	**56.67%/0.71**
no	no	58.58%/0.57	58.88%/0.58	35.47%/1.06	35.68%/1.04	**47.15%/0.81**
no	yes	60.80%/0.54	57.67%/0.56	40.02%/0.97	39.18%/1.00	**49.4%/0.77**
average		**63.18%/0.52**	**62.11%/0.54**	**41.3%/1.0**	**41.18%/1.02**	

먼저, 표 3에서 overlap을 적용한 경우와 그렇지 않은 경우를 살펴보면, 세그먼트 기반 평가 방식에서 overlap 방식을 적용한 경우 평균 63.18%의 F-score와 0.52의 ER 값을 가지며 그렇지 않은 경우엔 평균 62.11%와 0.54의 값을 가짐으로써 overlap 방식의 학습 기법이 보다 효과적임을 확인할 수 있다. 이벤트 기반 평가방식에서도 큰 차이는 아니지만 overlap 방식이 근소하게 우수한 성능을 나타냄을 확인할 수 있다.

표 3에서는 배치정규화와 드랍아웃이 성능에 어떤 영향을 미치는지도 확인 할 수 있는데, 예상한대로 배치정규화와 드랍아웃을 모두 적용한 경우에 가장 좋은 성능을 보임을 알 수 있었다. 배치정규화만 적용하고 드랍아웃을 배제한 경우에는 다소 낮은 성능을 보였지만 그 차이가 크지 않아서 드랍아웃의 성능에 대한 영향은 그리 크지 않음을 확인할 수 있었다. 한편, 배치정규화를 적용하지 않은 경우에는 드랍아웃의 적용여부와 상관없이 성능이 상당히 저하되는 것을 확인 할 수 있었으며, 배치정규화와 드랍아웃을 동시에 적용하지 않을시 가장 저조한 성능을 보임을 확인할 수 있었다.

표 4 학습 수렴 기준에 따른 CRNN 성능
Table 4. Performances of CRNN depending on the criterion of training convergence

criterion	segment-based	event-based	
	F-score/ER	F-score/ER	avergae
ER	67.24%/0.49	45.93%/0.94	**56.59%/0.72**
F-score	66.45%/0.51	44.97%/0.98	**55.71%/0.75**

표 4에는 학습수렴기준을 F-score로 한 경우의 성능을 ER 기준으로 한 경우와 비교하여 나타내었다. 배치정규화와 드랍아웃을 모두 적용하였으며 그 밖의 조건은 표 3과 동일하게 두었다. 전체적으로 ER 기준으로 한 경우가 F-score기준에 비해서 다소 나은 성능을 보임을 알 수 있으나 뚜렷하게 큰 차이는 없는 것으로 보아 CRNN의 학습에서 수렴기준으로는 두 가지 모두 채택 가능하다고 판단된다.

마지막으로, 표 5에는 CRNN의 입력단위인 세그먼트의 길이에 따른 성능 변화를 나타내었다. 세그먼트의 길이가 짧은 경우(2.56초, 5.12초)에 비해서 다소 긴 경우(10.24초, 20.56초)에 보다 우수한 성능이 나타남을 알 수 있다. 10.24초와 20.56초에서 비교적 좋은 성능이 나타나는 것은 본 연구에서 검출하고자 하는 소리 이벤트들 중에서 길이가 긴 것이 많은 것과 연관이 깊으며, 또한 RNN에서 비교적 긴 시간 관계 정보를 잘 탐지 할 수 있기 때문인 것으로 판단된다.

표 5 세그먼트 길이에 따른 CRNN 성능
Table 5. Performances of CRNN depending on the length of the segment

segment length(s)	segment-based	event-based	
	F-score/ER	F-score/ER	average
2.56	66.84%/0.51	42.75%/1.13	**54.80%/0.82**
5.12	67.47%/0.50	44.27%/1.04	**55.87%/0.77**
10.24	68.01%/0.47	45.33%/0.97	**56.67%/0.72**
20.56	67.37%/0.48	45.71%/0.92	**56.54%/0.70**

IV. 결 론

소리 이벤트 검출을 위한 다양한 기법 중에서 최근 주목을 받고 있는 방법은 딥러닝 기반의 CRNN 방식이다. 본 연구에서는 CRNN의 학습 과정에서 다양한 조건을 가정하고 인식 실험을 통하여 최적의 학습 방법을 찾고자 하였다.

CRNN에 대한 학습률은 10^{-4}로 설정하였을 때 가장 좋은 성능을 보임을 알 수 있었으며 이는 일반적으로 딥러닝 학습 과정에서 많이 사용하는 10^{-3}에 비해서 더 나은 성능을 보여줌을 인식실험을 통해서 확인 할 수 있었다. 또한 검증데이터에서 최고의 성능을 나타내는 학습률이 테스트 데이터에 대해서도 가장 좋은 성능을 보임을 알 수 있었으며, 이는 새로운 데이터나 딥뉴럴네트워크의 구조의 변화가 생길 경우, 검증데이터에 대한 성능 실험을 통해서 최적의 학습률을 설정하는 것이 적절함을 의미한다.

다른 딥 뉴럴네트워크와 마찬가지로 배치정규화와 드랍아웃은 CRNN의 성능을 향상시키는데 기여한다는 것을 확인 할 수 있었다. 특히, 배치정규화는 성능 향상에 대한 기여가 뚜렷함을 알 수 있었으며, 드랍아웃의 경우는 약간의 성능향상만을 나타내었다. 또한, 학습데이터를 매 반복 마다 동일하게 하는 것 보다는 각 세그먼트의 위치를 매 반복 마

다 변경함으로써 성능의 향상을 이룰 수 있음을 확인할 수 있었다.

소리 이벤트 검출에서 학습의 수렴 조건으로는 F-score와 ER을 많이 사용하는데 본 연구의 실험결과 ER을 사용할 경우 F-score에 비해서 약간의 성능 향상이 이루어짐을 확인 할 수 있었다. 마지막으로 CRNN의 입력 세그먼트의 길이를 다양하게 변환시킴으로써 성능 변화를 관찰 하였다. 세그먼트의 길이는 너무 짧은 경우보다는 다소 긴 경우가 더 나은 성능을 보임을 알 수 있었으며, 세그먼트의 길이가 10.24초와 20.56초에서 가장 좋은 성능을 얻었다.

References

[1] M. K. Nandwana, A. Ziaei, and J. H. L. 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.

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

[3] J. Salamon and J. P. Bello, "Feature Learning with Deep Scattering for Urban Sound Analysis", 23rd European Signal Processing Conference (EUSIPCO), pp. 724-728, Sep. 2015.

[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), pp. 2742-2746, Mar. 2016

[5] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic Classification of Multiple Simultaneous Bird Species: A Multi-instance Multi-Label Approach", The Journal of the Acoustical Society of America, Vol. 131, No. 6, pp.4640-4640, Jun. 2012.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems, pp. 1097-1105, Dec. 2012.

[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), pp. 6645-6649, May 2013.

[8] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. bengio, "Learning Phrase Representations Using Rnn Encoder-Decoder for Statistical Machine Translation", Proceedings of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724-1734, Oct. 2014.

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

[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. 25. No. 6, pp. 1291-1303, Jun. 2017.

[11] T. N. 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), pp. 4580-4584, Apr. 2015.

[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), pp. 2392-2396, March, 2017.

[13] TUT-SED Synthetic 2016,

http://www.cs.tut.fi/sgn/arg/taslp2017-crnn-sed/tut-sed-synthetic-2016 [accessed Sep. 01. 2018].

[14] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection", Applied Sciences, Vol. 6, No. 6, pp. 162-178, May 2016.

저자소개

정 석 환 (Suk-Hwan Chung)

2017년 3월 ~ 현재 : 계명대학교 전자공학과 대학원 석사과정

관심분야 : 머신러닝 및 오디오 분류

정 용 주 (Yong-Joo Chung)

1995년 8월 : 한국과학기술원 (공학박사)

1999년 3월 ~ 현재 : 계명대학교 전자공학과 교수

관심분야 : 오디오 분류, 음성인식