청각장애인용 자막방송 서비스를 위한 연쇄잔차 신경망 기반 음향 사건 분류 기법

김남균, 박동건, 김준호, 김홍국, *안충현
광주과학기술원 *한국전자통신연구원
{skarbs001, dongkeun, junhokim, hongkook}@gist.ac.kr * hyun@etri.re.kr

Sound Event Classification Based on Concatenated Residual Network Applicable to Closed Captioning Services for the Hearing Impaired

Nam Kyun Kim, Dong Keun Park, Jun Ho Kim, Hong Kook Kim, *Chung Hyun Ahn
Gwangju Institute of Science and Technology
* Electronics and Telecommunications Research Institute

요 약

본 논문에서는 청각장애인에게 자막방송을 제공하기 위하여 오디오 콘텐츠에 등장하는 음향 사건을 분류하는 기법을 제안한다. 제안된 기법은 복수의 잔차 신경망(ResNet)을 연결하는 연쇄잔차(concatenated residual) 신경망 구조를 갖는다. 신경망의 입력 특징을 위해 음성의 멜-주파수 켑스트럼 벡터를 다수의 프레임으로 결합하여 형성한 2 차원 이미지와 전체 프레임에 대한 멜-주파수 켑스트럼 벡터들로부터 얻은 1 차원의 통계 특징벡터를 얻는다. 각각의 입력은 2 차원 잔차 신경망과 1 차원 잔차 신경망으로 모델링되고, 두 개의 잔차 신경망을 연쇄연결(concatenation)하는 구조를 가진 연쇄잔차 신경망으로 구성된다. 성능평가를 위해 수집된 데이터 셋으로부터 6-fold 교차검증을 통해 평가한 결과, 85.48%의 분류 정확도를 얻을 수 있었다.

1. 서론

멀티미디어 콘텐츠는 텔레비전 방송, 인터넷 동영상 공유 플랫폼, 소셜 네트워크 서비스 등의 다양한 매체로 매 시간마다 생성되고 있다. 이렇게 생성된 콘텐츠들은 보는 사람에게 즉각적 반응을 요구하고 시간적 제약을 받는 환경에서 콘텐츠 이해를 돕기 위한 자막의 사용이 늘고 있다[1]. 그리고, 음성 자막뿐 아니라, 실감나는 방송 매체 제작을 위하여 생동감 있는 자막방송을 위한 연구가 진행 중이다[2]. 특히 청각장애인의 방송 이용확대 및 방송접근권 향상을 위해 제도적으로 의무화되는 추세이다[3]. 하지만, 이러한 자막 방송 제작을 위해서는 직접 듣고 이해하여 콘텐츠의 상황 이해 위해선 많은 시간과 노력을 필요로 한다. 따라서, 자동 자막 생성을 위한 머신러닝 기반의 방송 인지 기술들의 개발이 요구되고 있다.

이러한 콘텐츠에 포함된 상황을 이해하기 위해 음향 사건 관련 연구 그룹인 DCASE (Detection and Classification of Acoustic Scenes and Event) 챌린지가 매년 개최되어 음향 데이터를 활용한 상황인지 알고리즘 연구가 활발히 진행되고 있다. 종래의 연구로는 은닉 마르코프 모델, 서포트 벡터 머신, 비음수 행렬 분해와 같은 머신러닝 기법들을 활용하여 음향 사건 분류 기법들에 대한 연구가 보고 되었다[4]. 하지만 최근 들어 컴퓨팅 자원의 발전으로 음향 데이터 기반의 상황인지와 관련된 연구로는 음향 장면 분류, 음악 장르 분류 및 음향 사건 감지와 같은 문제에 대한 깊은 신경망 기반 접근방법들이 활용되었다. 특히 DCASE 2019 Challenge 에서는 합성곱 신경망 및 합성 순환 신경망 기반의 음향 사건 분류 기법이 제안되었다. 또한 잔차 신경망(ResNet)[5]은 비전과 오디오 분야에서 높은 정확도를 달성하는 것으로 나타났으며, 이는 각 층 간 skip connection 을 이용하는 잔차 학습을 활용한 음향 사건 분류 기법이 연구 중에

표 1. 음향 사건별 데이터 수 분포

사건 명	데이터 수	사건 명	데이터 수
toilet_flush	1,231	tire_skid	508
Water	1,164	thunder_stom	992
car_crash	1,713	bell	2,018
applaud	1,493	baby	744
knock	1,051	door	1,069
explosion	3,965	animal	2,411
siren	601	unknown	1,747
car_horn	2,171		

있다[6].

본 논문에서는 연쇄잔차 신경망을 활용한 음향 사건 분류 기법을 제안한다. 제안된 연쇄잔차 신경망의 구조는 2 차원 이미지와 1 차원 통계 특징을 입력으로 하는 2 차원 잔차 신경망과 1 차원 잔차 신경망을 연결하는 구조를 가진 연쇄잔차 네트워크와 같은 형태로 구성된다. 특히 제안된 신경망 구조의 음향 특징으로는 멜-주파수 켑스트럼(MFCC: mel-frequency cepstral coefficient)를 사용한다. 추출된 MFCC 에서 일정 프레임을 이미지 형태로 표현하여 2 차원 이미지 특징을 구성하고, 1 차원 특징은 추출된 이미지로부터 통계 특징들을 추출하여 활용한다. 제안된 신경망은 수집한 음향 데이터셋을 활용하였고, 이를 6 개 fold 로 분리하여 교차검증을 통해 성능 평가를 수행하였다.

본 논문의 구성은 다음과 같다. 2 절에서는 음향 사건 분류 데이터셋에 대해 설명한 후, 3 절에서는 본 논문에서 제안하는 연쇄잔차 신경망 구조에 대해 설명한다. 4 절에서는 제안된 기법의 성능 평가에 대해 논의한다. 마지막으로 5 절에서는 본 논문에 대한 결론을 맺는다.

2. 음향 사건 분류 데이터셋

본 연구에서 활용된 음향 사건 분류 알고리즘 개발을 위한 데이터셋은 DCASE 2018 챌린지의 데이터베이스 구성을 참고하여 구성하였고, 데이터셋 수집은 주로 오디오 셋[7] 에서 수집하여 정제하였다. 정의한 사건은 총 14 가지로 여기에 속하지 않는 미확인음을 포함하여 데이터셋을 구성하였다. 데이터셋에서 정의된 사건별 데이터수는 표 1 과 같다.

3. 연쇄잔차 신경망 기반 음향 사건 분류

먼저, 음향 사건 분류를 위해 오디오 신호는 16kHz 로

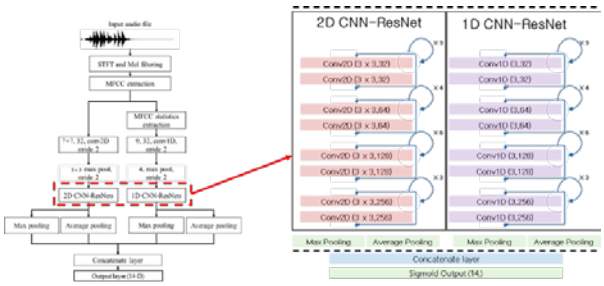

그림 1. 음향 사건 분류를 위한 제안된 연쇄잔차 신경망의 네트워크 구조

표 2. 입력 프레임 변화에 따른 분류 6 fold 교차검증 평균 정확도 비교

입력 프레임 수	1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	평균
200	83.64%	83.09%	83.82%	85.20%	85.36%	85.32%	84.41%
300	82.98%	85.01%	84.55%	85.68%	84.65%	83.27%	84.78%
500	85.76%	86.00%	84.97%	85.89%	85.49%	84.74%	85.48%

표본화되었으며, 10ms(= 160 샘플) 마다 이동하면서 30ms 해밍 윈도우를 적용한 후, 512 차 푸리에 변환을 적용한다. 그리고 음향 특징으로는 40 차의 MFCC 를 추출하였다. 여기에 1 차 차분과 2 차 차분을 취하여 총 120 차의 특징벡터를 얻었다. 각 10ms 마다 추출된 120 차 MFCC 특징벡터는 500 개(= 5sec) 프레임을 결합하여 2 차원의 이미지 형태로 표현하였다. 한편, 120 차의 MFCC 특징벡터를 각각의 차수마다 500 개 프레임에서의 평균, 표준편차, 왜도(skewness), 중앙값(median) 등 4 가지를 추출하여 총 480(= 120×4)차수를 갖는 1 차원의 특징벡터를 구하였다.

그림 1 은 본 논문에서 개발된 연쇄잔차 합성곱 신경망 기반 음향 사건 분류 모델의 구성도를 보여준다. 모델의 입력 특징들은 각각 2D 및 1D 잔차 신경망으로 입력한 후, 두 신경망이 결합되어 하나의 신경망으로 되는 구조로 구성된다. 우선, 120×500 차원의 이미지는 7×7 2D 합성곱 필터 32 개로 구성된 2 차원 합성곱 층을 통과한 후, 2 차 잔차 신경망(2D CNN-ResNet)에 인가된다. 2 차 잔차 신경망의 첫 번째 잔차 블록은 3 개의 잔차 유닛으로 구성되며 이는 3×3 의 합성곱 커널로 구성된다. 그리고 2, 3, 4 번째의 잔차 블록은 각각 4, 6, 3 개의 residual unit 으로 구성되어 신경망을 형성하며, 각각의 합성곱 커널의 수는 64, 128, 256 개로 구성되고, 각각의 합성곱 커널은 rectified linear unit (ReLU) 활성함수를 활용하여 훈련된다.

다음으로, 480 차의 1 차원 벡터 형태의 입력특징은 9×1 1D 합성곱 필터 32 개를 갖는 1 차원 합성곱 층을 통과한 후, 1 차원 잔차 신경망 (1D CNN-ResNet)에 인가된다. 1D ResNet 의 잔차 블록의 첫 번째 잔차 블록은 3 개의 잔차 유닛으로 구성되며 이는 3×1 의 합성곱 커널로 구성된다. 그리고 2, 3, 4 번째의 잔차

블록은 각각 4, 6, 3 개의 잔차 유닛으로 구성되어 신경망을 형성한다. 그리고 각각의 합성곱 커널의 수는 64, 128, 256 개로 구성되고, 각각의 합성곱 커널은 또한 ReLU 활성함수를 활용하여 훈련되었다.

최종적으로 2D ResNet 과 1D ResNet 의 출력 특징맵의 노드 값을 pooling 하는 함수로 각각 global max pooling 과 global average pooling 을 적용하여 각각 2 개의 특징 벡터를 생성하고, concatenate layer 를 통해 총 4 개의 특징벡터를 결합한다. 이렇게 결합된 특징벡터는 fully connected layer 를 통해 출력벡터와 연결된다. 출력벡터는 사전에 정의한 음향 사건 개수인 14 개로 정의하였고, 활성함수는 sigmoid 가 활용되었다.

훈련을 위해서 미니 배치(minibatch)의 크기는 256, 학습률(learning rate)은 0.001, 최적화 함수로는 Adam 알고리즘[8]을 활용하였으며, 최대 80 epochs 동안 훈련하여 교차검증 시 활용된 데이터의 validation loss 가 제일 낮은 모델을 저장하였고, 손실값(loss) 개선이 이루어지지 않을 때 학습을 멈추도록 하였다. 각 신경망의 학습 및 평가는 Tensorflow 를 백엔드로 사용하는 Keras 프레임 워크를 활용하여 구현되었다.

4. 성능 평가

본 논문에서 제안된 음향 사건 모델의 성능 평가를 위해 2 절에서와 같이 구축된 DB 를 활용하였고, 최종적으로 활용된 총 22,878 개의 음향 데이터로 구성되어 있으며, 6-fold 교차검증(cross-validation)을 통해 모델 성능평가를 진행하였다. 본 실험은 분류 정확도(accuracy)를 성능 지표로 하였다. 합성곱 신경망은 2 차원 신호로 변환된 오디오 신호를 입력으로 사용할 경우 음향 신호의 시간적인 특징 및 주파수 특성을 동시에 모델링 할 수 있다. 이때 몇 개의 시간 프레임을 하나의 입력 이미지로 구성하는 지에 따라 합성곱 신경망으로 입력되는 이미지의 크기가 달라지며, 이에 따른 성능평가를 수행하였다.

<표 2>는 입력 프레임 수에 따른 음향 사건 분류의 교차검증 정확도를 보여 준다. 표에서 보는 바와 같이, 본 과제에서 개발된 음향 사건 분류 모델의 경우 전체 정확도 비교 시, 500 프레임을 활용한 경우 교차검증 평균 정확도 85.48%를 얻을 수 있었고, 시간 프레임이 커질수록 정확도가 높아짐을 확인할 수 있었다.

5. 결론

본 논문에서는 음향 사건 분류를 위해 연쇄잔차 신경망을 제안하였다. 음향 사건 분류를 위해 구성한 데이터셋은 DCASE 2018 챌린지에 활용된 데이터셋 구성 및 규격을 참고하여 수집하였다. 제안된 연쇄잔차 신경망 기반의 음향 사건 분류를 구축된 데이터셋을 교차검증을 통해 그 성능을 평가하였다. 평가 결과, 입력 오디오 신호의 500 프레임을 활용한 음향 사건 분류 결과의 정확도가 85.46%를 얻을 수 있었고, 입력 이미지의 크기가 증가함에 따라 정확도가 상승하는 것을 확인할 수 있었다.

감사의 글

본 연구 논문은 과학기술정보통신부 및 정보통신기획평가원의 출연금으로 수행하고 있는 한국전자통신연구원 시청각 장애인의 방송시청을 지원하는 감성표현 서비스 개발[2019-0-00447]의 연구결과입니다.

참고문헌

[1] 정수영, “TV 영상자막의 특징 및 기능에 관한 연구: 지상파 TV 3 사의 리얼 버라이어티쇼를 중심으로,” 한국언론학보, 제 53 권, 제 6 호, pp. 153-176, 2009.

[2] 김성훈, 안충현, 서봉석, 현창중, 김동호. “AT SC3.0 기반 청각장애인을 위한 감정표현자막방송 핵심기술에 관한 연구,” 대한전자공학회 학술대회, pp. 233-234, 2019.

[3] 안충현, “장애인방송 기술개발 현황,” 전자자통신동향분석, 제 34 권, 제 3 호, 2019.

[4] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, “Sound event detection in the DCASE 2017 Challenge,” IEEE/ACM Trans. Audio Speech Language Processing, vol. 27, no. 6, pp. 992-1006, June 2019.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[6] Y. Kiyokawa, S. Mishima, T. Toizumi, K. Sagi, R. Kondo, and T. Nomura, “Sound event detection with ResNet and self-mask module for DCASE 2019 Task 4,” Technical Report, June 2019.

[7] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. 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, pp 776-780, 2017.

[8] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in Proc. of the 3rd International Conference on Learning Representations, 2014.