자막방송을 위한 잔차 합성곱 순환 신경망 기반 음향 사건 분류

김남균, 김홍국, *안충현
광주과학기술원 *한국전자통신연구원
{skarbs00, hongkook}@gist.ac.kr *hyun@etri.re.kr

Residual Convolutional Recurrent Neural Network-Based Sound Event Classification Applicable to Broadcast Captioning Services

Nam Kyun Kim, Hong Kook Kim, *Chung Hyun Ahn
Gwangju Institute of Science and Technology
* Electronics and Telecommunications Research Institute

요약

본 논문에서는 자막방송 제공을 위해 방송콘텐츠를 이해하는 방법으로 잔차 합성곱 순환신경망 기반 음향 사건 분류 기법을 제안한다. 제안된 기법은 잔차 합성곱 신경망과 순환 신경망을 연결한 구조를 갖는다. 신경망의 입력 특징으로는 멜-필터뱅크 특징을 활용하고, 잔차 합성곱 신경망은 하나의 스템 블록과 5개의 잔차 합성곱 신경망으로 구성된다. 잔차 합성곱 신경망은 잔차 학습으로 구성된 합성곱 신경망과 기존의 합성곱 신경망 대비 특징맵의 표현 능력 향상을 위해 합성곱 블록 주의 모듈로 구성한다. 추출된 특징맵은 순환 신경망에 연결되고, 최종적으로 음향 사건 종류와 시간정보를 추출하는 완전연결층으로 연결되는 구조를 활용한다. 제안된 모델 훈련을 위해 라벨링되지 않는 데이터 활용이 가능한 평균 교사 모델을 기반으로 훈련하였다. 제안된 모델의 성능평가를 위해 DCASE 2020 챌린지 Task 4 데이터 셋을 활용하였으며, 성능 평가 결과 46.8%의 이벤트 단위의 F1-score를 얻을 수 있었다.

1. 서론

일반적으로 자막방송은 방송콘텐츠에 별도 자막을 부가방송 형태로 제공하는 것으로, 이러한 자막방송은 어린이나 외국인들의 어학학습을 위해 이용되거나 청각장애인을 위한 방송해설을 위해 제공된다 [1]. 자막의 종류는 화면에 속기를 통해 실시간으로 제공되는 폐쇄 자막과 감정표현 등과 같은 화면해설 이외에 감성자막표현을 위한 개량형 자막이 있다. 특히 개량형 자막은 수도권 9개 복지관을 대상으로 선호도 조사 결과, 기존 폐쇄 자막 대비 감정적 표현이 추가되어 선호도가 높음을 보였다[2]. 하지만, 이러한 개량형 자막 생성을 위한 감성 정보 확보는 방송콘텐츠의 상황 이해를 위해 직접 듣고 생성해야 하므로, 자동 생성을 위한 연구가 필요로 된다.

콘텐츠 이해를 위한 방법으로는 음향 사건 분류 기법을 활용하여 콘텐츠 내 존재하는 음향 정보 생성이 필수적이다. 음향 사건 분류 기법으로는 서포트 벡터 머신, 은닉 마코프 모델 등과 같은 기계학습 기반의 모델이 주로 연구되었다. 최근 들어, 합성곱 신경망과 순환신경망, 그리고 합성곱 순환 신경망 모델 기반의 알고리즘이 제안되었으며[3], 평균 교사 모델을 활용하여 라벨링되지 않은 데이터를 활용하는 등 관련 연구가 활발히 진행되고 있다[4].

본 논문에서는 잔차 합성곱 순환 신경망(RCRNN: residual convolutional recurrent neural network)을 활용한 음향 사건 분류 기법을 제안한다. 제안된 모델은 잔차 합성곱 신경망과 순환 신경망의 조합으로 구성된다. 특히 입력 음향 특징을 모델링하는 합성곱 신경망은 레이어 수를 많이 활용할수록 성능 향상을 기대할 수 있다. 이 경우, 기울기 소실 문제를 야기할 수 있으므로 제안된 모델에서는 잔차 학습을 통해 이를 극복한다. 또한 합성곱 순환 신경망으로 추출된 특징맵의 표현능력 향상을 위해 합성곱 블록 주의 모듈(CBAM: convolutional block attention module)[5]을 적용한다.

표 1. DCASE 2020 챌린지 Task 4 데이터 셋 구성

Dataset		Number of clips
Training set	Strongly labeled dataset	2,584 clips
	Weakly labeled dataset	1,578 clips
	Unlabeled dataset	14,412 clips
Development set	Validation test dataset	1,168 clips

2. DCASE 2020 챌린지 Task 4 데이터 셋

본 연구에서의 알고리즘 개발 및 성능 평가는 <표 1>과 같이 DCASE 2020 챌린지 Task 4 데이터 셋을 활용한다. 표에서 보는 바와 같이, 데이터 셋은 크게 강하게 라벨링된(strongly labeled) 데이터 셋, 약하게 라벨링된(weakly labeled) 데이터 셋 그리고 라벨링되지 않은(unlabeled) 데이터 셋을 활용하여 모델을 훈련하고, development 셋으로 제공된 데이터로 성능을 평가한다. 본 데이터 셋은 10가지 음향 사건 즉, speech, dog(barking), cat, alarm/bell/ringing, dishes, frying, blender, running water, vacuum cleaner, electric shaver로 구성된다.

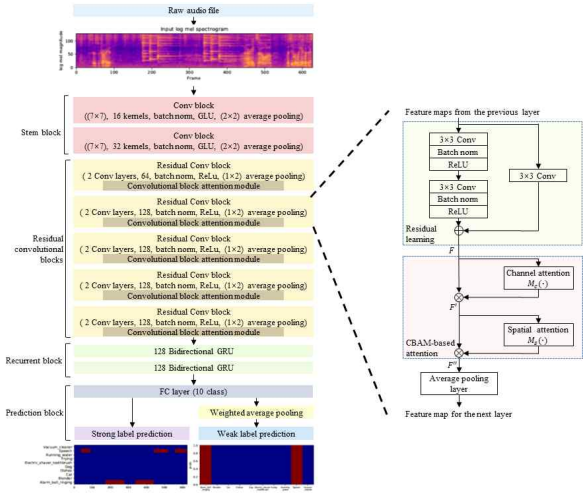

그림 1. 제안된 잔차 합성곱 순환 신경망(RCRNN) 구조

3. 잔차 합성곱 순환 신경망 기반 음향 사건 분류

<그림 1>은 제안된 CBAM이 적용된 RCRNN 구조를 보여 준다. 먼저 [4]에 설명된 것과 같이, 주어진 오디오 클립의 신호는 44.1 kHz에서 16 kHz로 다운샘플링되고, 255 샘플 홉 크기로 2,048개 샘플 마다 하나의 프레임으로 분할된다. 그리고 나서, 2,048-포인트 고속 푸리에 변환을 적용하여 128차원의 로그 멜-필터뱅크(log mel-filterbank)로 변환된다. 다음으로, 10초 분량의 프레임인 628개의 프레임을 그룹화하여 (628×128) 차원의 스펙트럼 이미지를 만든 다음 RCRNN의 입력 특징으로 사용한다.

그림에서 보는 바와 같이, RCRNN은 하나의 스템 블록과 5개의 잔차 합성곱 블록으로 구성되며, 스템 블록은 1, 2 번째 합성곱 블록에 대해 각각 16, 32개의 (7×7) 합성곱 필터, 배치 정규화, 게이트 선형 유닛(GLU: gated linear unit) 활성화 및 (2×2) 평균 풀링층으로 순차적으로 구성된다. 잔차 합성곱 블록은 각각 32, 64, 64, 128, 128개의 (3×3) 합성곱 필터로 구성하고, 배치정규화 그리고 정류 선형 유닛 (ReLU: rectified linear unit)으로 구성되며 이는 CBAM으로 연결되어 잔차 합성곱 블록을 구성한다. 다음으로, 잔차 합성곱 블록으로부터 추출된 특징맵은 순환 신경망 블록에 연결된다. 순환 신경망 블록은 두 개의 양방향 게이트 순환 유닛(BiGRU: bidirectional gated recurrent unit)으로 구성되어 추출된 특징맵의 시간적 컨텍스트 정보를 학습한다. 순환 신경망 블록의 (157×256) 출력은 완전연결층에 의해 처리된 다음 시그모이드 함수에 의해 처리되어 (157×10) 차원의 출력이 생성된다. 여기서 10은 분류할 음향 사건 수를 나타낸다. 결과적으로 628의 입력 차원은 157의 출력 차원으로 축소되고, (157×10) 차원 출력은 음향사건 분류 유형 및 시간 정보를 포함하는 강력한 라벨(strong label)을 예측한다.

모델 훈련을 위해 본 논문에서는 레이블이 없는 데이터를 사용할 수 있는 학습 방법으로 평균 교사 기법 및 을 사용하였다. 평균 제곱 오차 함수를 교사 모델과 학생 모델의 일관성 유지를 위해 사용하고 학습률을 ramp up으로 증가하였고 일관성 비용을 2.0으로 설정하였다[4]. 마지막으로 데이터 증강으로 시간 및 주파수 시프팅을 적용하였다 [6].

표 2. DCASE 2020 챌린지 Task 4에 적용된 모델별 성능 비교

Model	Event-based F1-score
Baseline of DCASE 2020 Task 4 [4]	34.8
Top-ranked model of DCASE 2020 Task 4 (single model) [6]	46.0
Proposed RCRNN model	46.8

4. 성능평가

본 논문에서 제안된 음향 사건 분류 모델의 성능을 평가하기 위해 2절에서 설명한 바와 같이 DCASE 2020 챌린지 Task 4 데이터 셋을 활용하였다. 모델 비교를 위해 합성곱 순환 신경망 기반의 DCASE 2020 챌린지 Task 4 baseline 모델 및 해당 챌린지 1위 모델과, 제안된 RCRNN 모델의 성능비교를 진행하였다. 본 실험은 이벤트 단위의 F1-score를 성능 지표로 하였다.

<표 2>는 모델별 성능을 비교하여 보여준다. 표에서 보는 바와 같이, 본 논문에서 제안된 RCRNN 기반 음향 사건 분류 모델은 baseline과 챌린지 1위 모델 대비, 23.0%와 0.8% 높은 F1-score를 보였다.

5. 결론

본 논문에서는 음향 사건 분류를 위한 RCRNN을 제안하였다. 성능 평가를 위해 DCASE 2020 챌린지 Task 4 데이터 셋를 활용하여 각각의 이벤트 단위의 F1-score 성능을 비교하였다. 평가 결과, 챌린지 baseline 모델 대비 23.0% 높은 F1-score를 얻을 수 있었으며, 특히 해당 챌린지 1위 모델 대비 0.8% 높은 F1-score를 얻을 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 연구결과임.

참고문헌

[1] 송종길, “장애인의 방송접근권 확대를 위한 정책방안 연구,” 방송연구, 제57호, pp.147-178, 2003.
[2] 안충현, “장애인방송 기술개발 현황,” 전자통신동향분석, 제34권, 제3호, 2019.
[3] 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 Lang. Process., vol. 27, no. 6, pp. 992-1006, June 2019.
[4] N. Turpault and R. Serizel, “Training sound event detection on a heterogeneous dataset,” in Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), pp. 200-204, 2020.
[5] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. European Conference on Computer Vision (ECCV), pp. 3-19, 2018.
[6] 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.,” in Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), pp. 200-204, 2020.