음향 장면 분류에서 히트맵 청취 분석

*서상원, **박수영, ***정영호, ****이태진
전자통신연구원
{*suhsw1210, **sooyoung, ***yhcheong, ****tjlee}@etri.re.kr

Listenable Explanation for Heatmap in Acoustic Scene Classification

*Sangwon Suh **Sooyoung Park ***Youngho Jeong ****Taejin Lee
Electronics and Telecommunications Research Institute

요 약

인공신경망의 예측 결과에 대한 원인을 분석하는 것은 모델을 신뢰하기 위해 필요한 작업이다. 이에 컴퓨터 비전 분야에서는 돌출맵 또는 히트맵의 형태로 모델이 어떤 내용을 근거로 예측했는지 시각화 하는 모델 해석 방법들이 제안되었다. 하지만 오디오 분야에서는 스펙트로그램 상의 시각적 해석이 직관적이지 않으며, 실제 어떤 소리를 근거로 판단했는지 이해하기 어렵다. 따라서 본 연구에서는 히트맵의 청취 분석 시스템을 제안하고, 이를 활용한 음향 장면 분류 모델의 히트맵 청취 분석 실험을 진행하여 인공신경망의 예측 결과에 대해 사람이 이해할 수 있는 설명을 제공할 수 있는지 확인한다.

1. 서론

컨벌루션 신경망 (Convolution Neural Network, CNN)은 인간의 시각 인지 프로세스를 모델링한 신경망 구조로써, 이미지 분류, 객체 탐색 등 컴퓨터 비전 분야의 다양한 문제들을 풀고 있다. 그 구조는 수백만에서 수천만 파라미터 이상으로 구성되어 있고, 따라서 신경망의 기작을 해석하는 것은 복잡한 작업이다. 하지만, 모델의 예측 근거를 알 수 없다면, 자율 주행이나 의료 진단 등 높은 신뢰도가 요구되는 분야에서 활용이 곤란하다. 또한 일부 국가에서는 알고리즘에 의해 결정된 사안은 그 이유도 함께 제공하도록 규정하고 있기에[1], 신경망 모델의 예측 결과를 해석할 수 있어야 해당 시장에 서비스가 가능하다.

이미지 도메인에서는 인공신경망의 예측 근거를 입력 이미지 상에 시각화 하는 해석 방법이 연구되었다. 2013 년 Simonyan [2] 등은 모델의 분류 점수에서 역전파(backpropagation)되는 구배(gradient)를 활용하여 클래스 별 돌출맵(saliency map)을 구성하는 방법을 제안했다.

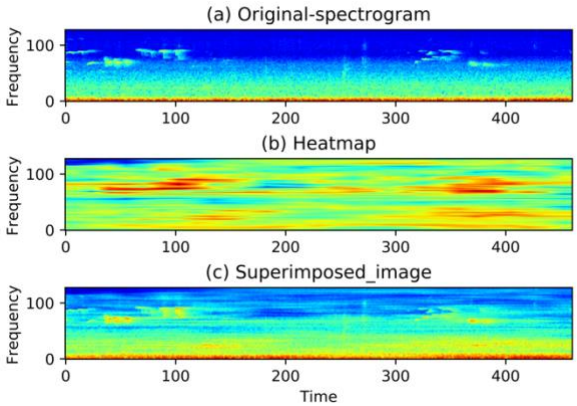

그림 1 (a)스펙트로그램, (b)히트맵, (c)Superimposed 이미지

이 방법은 이미지 속의 객체의 윤곽과 질감에 대해 시각화된 근거를 제시했으며, 이는 2014 년에 Zeiler [3]와 Springenberg [4]의 연구로 이어진다. 하지만 2018 년 Abedayo [5]의 연구에서 Springenberg 의 Guided Backpropagation 이 무결성 검사(sanity check)를 통과하지 못하였고, 같은 이유로 Zeiler 의

방법 또한 신뢰성에 의문이 있는 상황이다. 앞서 언급한 무결성 검사에서 유효한 것으로는 Simonyan 의 방법 외에도 Gradient-weighted Class Activation Mapping (Grad-CAM) [6]이 있다. Grad-CAM 은 Class Activation Mapping (CAM) [7]의 일반화된 방법으로 모델의 순전파(forward propagation) 과정에서 생성되는 피쳐맵(feature map)을 기반으로 히트맵을 생성한다. 모델 내부의 피쳐맵은 풀링(pooling) 계층 등을 통과하며 이미지의 압축이 발생하므로, 이를 기반으로 생성한 히트맵은 입력 이미지 크기로 늘리는 과정이 추가로 필요하다.

오디오 도메인에서도 히트맵을 활용한 컨벌루션 신경망의 분석 사례가 있다. Yuzhong Wu [8] 등은 음향 장면 분류 모델을 해석하기 위해 CAM 과 Grad-CAM 으로 히트맵을 구성하였으며, 모델이 특징적인 음향 이벤트 보다는 음향 질감(texture)에 근거하여 예측을 수행한다고 보고했다. 하지만 그림 1 과 같이 오디오 스펙트로그램 상에 나타나는 히트맵은 그 정보가 무엇인지 직관적으로 해석하기 어렵다. 이를 확인하는 방법은 직접 모델의 예측에 영향을 준 소리를 듣고 판단하는 것이다.

본 연구에서는 오디오 도메인의 컨벌루션 모델 히트맵 분석을 위한 히트맵 청취 분석 시스템을 제안한다. 저자들은 2019 년도 DCASE 챌린지에 제출된 음향 장면 분류 모델을 대상으로 청취 분석 실험을 진행했으며, 장면 클래스와 관련 있는 소리가 검출된 샘플들을 확인할 수 있었다.

2. 배경

2.1. DCASE 음향 장면 분류

음향 장면 분류는 데이터를 분석하여 그것이 녹음된 장소 클래스 중 하나로 분류하는 작업으로, Detection and Classification of Acoustic Scenes and Events (DCASE) 챌린지의 주요 태스크로써 처음 개최된 2013 년부터 다뤄지고 있다. DCASE 챌린지에서는 매년 10 초 길이의 데이터(2016 년까지는 30 초)로 구성된 음향 장면 데이터세트를 업데이트하고 있고, 참가자들은 이를 활용하여 모델을 훈련할 수 있다. 음향 장면 분류 시스템의 평가 척도는 클래스 별 분류 정확도이며, 평가를 위한 데이터세트는 라벨 정보 없이 제공된다. 2018 년부터 음향 장면 분류 작업은 아래 세 가지 하위 작업으로 나뉜다.

- Subtask A: 고성능 바이노럴 마이크에서 녹음된 데이터를 활용한 장면 분류
- Subtask B: Subtask A 의 데이터에 모바일 기기 등 다양한 녹음 장치에서 수집된 데이터를 포함하는 음향 분류
- Subtask C: 외부 데이터를 활용한 모델 훈련을 허용하는 작업

매년 DCASE 에서는 참가자들이 어떤 방법으로 문제를 해결했는지 공개하고 있다. 공개된 기록에 따르면 2017 년 이후 대부분의 시스템에서 주요 입력 피쳐로는 로그-멜 스펙트로그램과 같은 주파수 도메인 데이터가, 분석 모델로는 컨벌루션 신경망이 활용되고 있다. 음향 장면 분류 작업에서 Open source Award 를 수상한 시스템 [9]은 각 하위 작업의 평가 데이터에서 80.5%, 74.9% 및 58.8%를 기록하였다.

2.2. Grad-CAM 히트맵 생성 기법

Grad-CAM 은 모델의 순전파 과정에서 생성되는 피쳐맵을 결합하여 히트맵을 구성하는 방법이다. 각 피쳐맵은 현재 클래스 분류 점수에 대해 가중치가 결정되며, 이 가중치는 역전파된 구배값의 평균으로 결정된다. $f_{k(x,y)}$ 가 히트맵을 구성할 계층의 k 번째 피쳐맵이라고 가정하자. 이때 클래스 c에 대한 Grad-CAM $M_c(x,y)$는 아래 식 (1)과 같이 정의할 수 있다.

$$M_c(x,y) = ReLU(\sum_k \alpha_k^c f_{k(x,y)}) \quad (1)$$

여기서 α_k^c 는 각 피쳐맵의 가중치이며, 분류 점수 y^c 의 그래디언트 평균이다. 그리고 생성된 히트맵에서 필요한 부분은 양성 활성화 (positive activation) 영역이므로 Rectified Linear Unit (ReLU) 함수를 적용하고 있다.

$$\alpha_k^c = \frac{1}{Z}\sum_{i,j}\frac{\partial y^c}{\partial f_k(i,j)} \quad (2)$$

Grad-CAM 으로 생성된 히트맵은 입력 데이터보다 작아진 상태이므로 이를 리사이즈 해야 하며, 선형보간법으로 일정하게 크기를 늘리는 방법이 일반적이다. 하지만 모델 안에서 이미지를 비선형으로 압축한 경우에는 이를 고려하여 크기를 보정해야 히트맵 상의 공간 정보를 유지할 수 있다.

3. 실험 방법

히트맵의 청취 분석 실험을 위해 히트맵 청취 분석 시스템을 만들었으며, 이를 활용하여 음향 장면 분류 모델의 히트맵 영역 소리를 추출하는 실험을 진행하였다.

3.1. 음향 장면 분류 모델

본 연구에서는 DCASE2019 챌린지 음향 장면 분류 작업에서 Open source Award 를 수상한 모델 [9]을 분석 대상으로 선정하였다 (그림 2). 이 모델은 소스 코드가 공개되어 있고 훈련된 모델 파라미터가 제공되어 재현 가능하며, 다중 모델의 앙상블 구조가 아니기에 히트맵 생성이 용이하여 본 실험에 가장 적절한 모델로 판단되었다. 모델의 입력은 128 개의 로그 멜 스펙트로그램과 이에 대한 delta 및 delta-delta 를 스테레오 채널 각각에 대해 생성하여 여섯 개의 이미지를 쌓은 구조로 되어 있다. 모델 내부에서의 피쳐맵 축소 과정은 시간 축에 대해서만 이루어지며, 주파수 축에 대해서는 고주파수 영역과 저주파수 영역으로 분할되여 병렬 경로로 처리된다. [9]의 저자는 이 분할 구조에서 저주파와 고주파에 대해 각각 다른 특징을 배울 수 있다고 설명한다. 각각의 병렬 경로는 ResNet 구조로 구성되었으며, 두 경로를 연결(concatenate)한 뒤 1x1 컨볼루션으로 분류 점수를 계산한다.

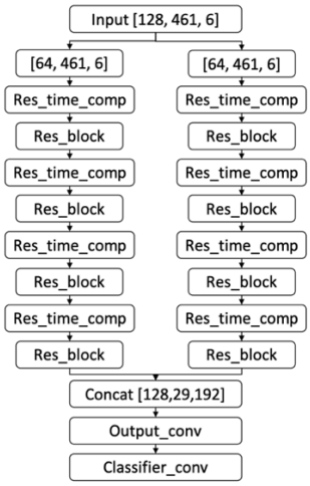

그림 2 주파수 축 병렬 경로 구조의 음향 장면 분류 모델

3.2. 데이터세트

본 실험에서는 신뢰할 수 있는 샘플을 다수 얻기 위해 세 개의 하위 작업 중에는 가장 분류 정확도가 높은 Subtask A 의 데이터세트를 대상으로 선정하였다. 3.1 절의 모델을 훈련하고 평가하는데 사용된 TAU Urban Acoustic Scene 2019 development dataset 은 유럽 주요 도시 10 곳에서 녹음된 10 개의 장면 클래스로 구성되며, 클래스 균등한 14,400 개의 파일이 포함되어 있다. 각 파일은 샘플레이트가 48 kHz 인 10 초 길이의 스테레오 오디오이다. 주최측에서는 9,185 개의 훈련 데이터와 4,185 개의 테스트 데이터를 포함하는 분할 메타데이터를 제공한다. 히트맵 분석의 목적은 훈련 단계에서 학습한 모델 내부 피쳐맵을 찾는 것이기 때문에, 모델이 훈련 과정에서 본 데이터를 다시 히트맵 생성에 활용하였다.

3.3. 히트맵 청취 분석 시스템

히트맵 청취 분석 시스템의 구조는 그림 3 와 같이 히트맵 마스크 생성 경로와 마스킹 신호 복원 경로로 구성되어 있다. 먼저 히트맵 마스크의 생성 과정을 살펴보면, 입력된 스테레오 오디오 신호는 푸리에 변환(FFT)되어 주파수 영역의 데이터(Mag/Phase data)로 변환된다. 2 채널의 주파수 영역 데이터는 피쳐 변환기(Feature extract)에서 로그 멜 스펙트로그램과 그의 전후 2 프레임에 대한 델타(delta) 및 전후 4 프레임에 대한 델타-델타(delta-delta) 피쳐로 변환되어 전체 6 채널의 오디오 입력 피쳐로 변환된다. 오디오 피쳐로 변환하는 과정에서 원분 데이터의 주파수 축에 대한 멜 스케일 압축이 발생하며, 이는 비선형의 변환이다. 따라서 히트맵의 리사이즈(Resize) 단계에서 이를 고려한 크기 복원을 할 수 있도록 크기 변형 정보(Size information)를 전달한다. 히트맵은 모델의 최상단에 위치한 컨벌루션 계층에서 생성된 피쳐맵들의 가중합으로 구성되었다. 각 피쳐맵의 가중치는 Grad-CAM 과 동일한 방식으로 클래스 분류 점수에서 역전파되는 구배의 평균을 사용한다. 본 저자들은 히트맵을 이진화(Binarize)하여 활성화 영역 밖의 소리를 모두 마스킹 했을 때 보다 청취 분석이 용이함을 확인하였다. 히트맵의 이진화 방법은 정규화 된 히트맵 M과 이진화 임계값 θ, 스케일 보정값 δ에 대해 아래 식 (3)과 같은 시그모이드(Sigmoid) 함수로 나타낼 수 있다. 본

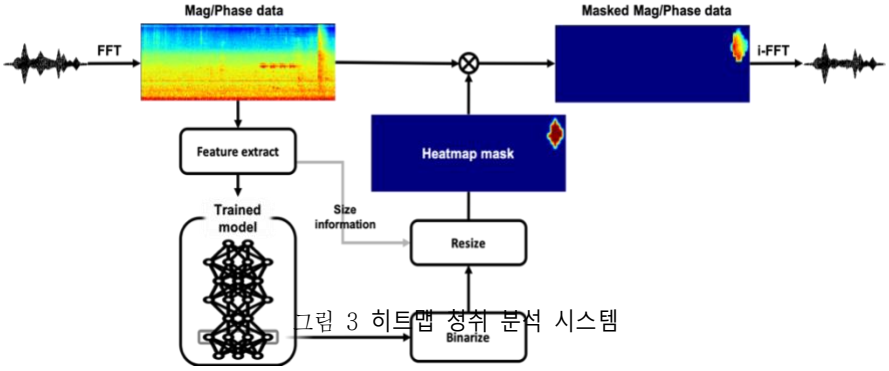

그림 3 히트맵 청취 분석 시스템

실험에서 임계값은 히트맵의 상위 25%의 활성화 값에 대해 청취하기 위해 75 백분위수에 해당하는 값으로 설정되었다.

$$M_{bin} = \frac{1}{1+\exp\left(-10^{\delta}(M-\theta)\right)} \quad (3)$$

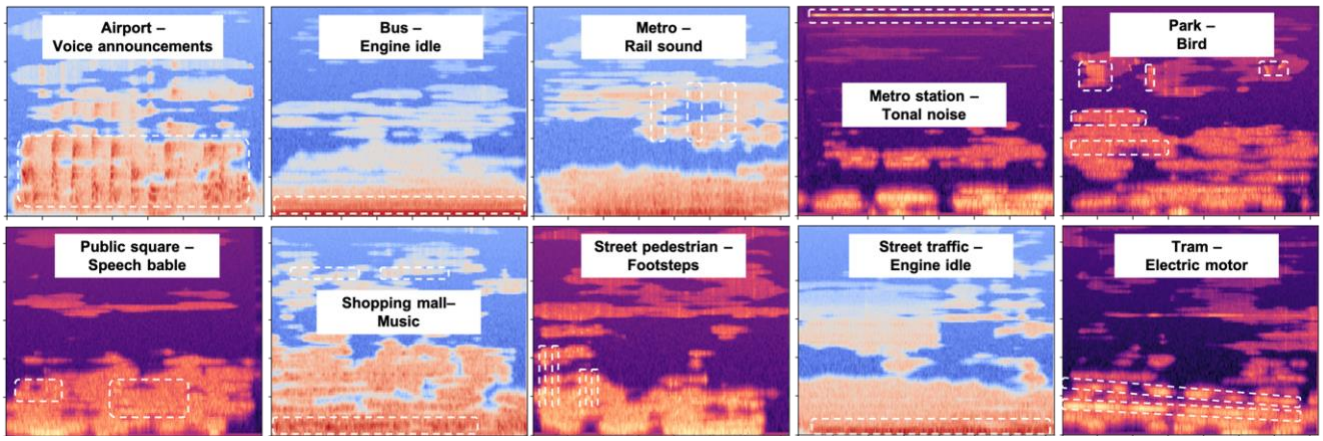

그림 4 음향 장면 별 검출된 음향 이벤트

마스킹 신호 복원 경로는 앞서 생성된 주파수 영역 데이터에서 시작한다. 히트맵 마스크는 주파수 영역 데이터와 같은 크기의 행렬이므로, 활성화 영역 밖의 정보가 마스킹 된 데이터(Masked Mag/Phase data)는 둘의 요소별(element-wise) 곱으로 표현할 수 있다. 이를 역 고속 푸리에 변환(inverse FFT)하여 시간 영역의 표현으로 바꾸면 활성화 영역에 대해 청취할 수 있는 오디오 신호를 얻을 수 있다.

4. 실험 결과

결과 분석에는 9,185 개의 훈련 데이터 중에서 소프트맥스(Softmax) 분류 점수 기준으로 0.8 점 이상이 나온 467 개의 데이터가 활용되었다. 이는 다른 클래스로 혼동한 정보가 많은 데이터를 본 실험 분석 대상에서 배제하고, 모델의 분류 결과에 강한 영향을 주는 소리를 찾기 위함이다.

마스킹된 데이터의 스펙트로그램 분석(그림 4)에서 우선 음향 이벤트로 보이는 정보들이 활성화 영역에 포함되 있는 것을 확인할 수 있었다. 활성화 영역은 몇몇 이벤트를 포함하며 시간 축에 대해 길게 늘어진 형상을 보이고 있는데, 이는 실험 대상 모델이 시간 축에 대해 모델 내부 피쳐맵을 축소한 것을 히트맵 생성을 위해 다시 늘리는 과정에서 본래의 활성화 영역보다 넓게 나타난 것이다.

분석 대상 데이터를 청취했을 때, 1/3 정도의 데이터에서 해석 가능한 소리를 확인할 수 있었다. 각 장면 별로 청취 분석이 가능했던 정보는 그림 4 의 흰색 사각형으로 표기했으며, 아래에 장면 별로 그 내용을 정리하였다. 청취 해석이 가능한 샘플들은 저자의 Github Pages[1]에서 확인할 수 있다.

- Airport: 공항의 안내 방송 목소리가 검출
- Bus: 버스의 엔진음이 검출되었으며, 다른 데이터에서는 차량의 흔들림이나 버스 문이 열리고 닫히는 소리 확인
- Metro: 철로 위를 달리는 소리가 검출되었으며, 다른 데이터에서는 객차 내 안내 방송음과 객차 문 닫히는 소리를 확인
- Metro station: 청취는 불가하나 고음의 톤 시그널 노이즈 발생 확인
- Park: 새 소리 검출
- Public square: 사람들의 웅성거림, 음악소리, 발자국 소리 등이 섞인 환경음 검출
- Shopping mall: 음악 소리가 포함된 넓은 환경음 검출
- Street pedestrian: 발자국 소리와 사람들의 웅성거림이 포함된 환경음 확인
- Street traffic: 엔진음, 차량 및 오토바이가 지나가는 소리 등을 검출
- Tram: 전동차의 모터 소리 검출

5. 결론

본 연구에서는 스펙트로그램 상의 히트맵을 청취하여

1 https://sangwonsuh.github.io/listenable_heatmap/

해석하는 방법을 제안한다. 이로써 기존의 시각적 분석으로는 해석하기 어려운 오디오 도메인의 히트맵을 청취함으로써 실제 모델 예측에 영향을 준 소리가 무엇인지 해석하는 방법을 제안한다. 저자들은 기 훈련된 음향 장면 분류 모델과 DCASE 음향 장면 데이터세트를 활용하여 히트맵을 생성하고 청취하는 실험을 진행하였다.

실험 결과에서 청취하여 해석할 수 있는 다수의 샘플을 확인하였으며, 그 중에는 음향 이벤트를 포함하고 있어 선행 연구[8]의 사례와 다른 점을 확인하였다. 이 실험을 통해 음향 장면 분류 모델이 스펙트로그램 상의 특징적인 이벤트를 학습할 수 있고, 이것이 장면 특징적인 정보라면 음향 장면 분류에 기여할 수 있음을 확인하였다.

본 연구에서 제안하는 히트맵의 청각 분석 방법은 컨볼루션 신경망 내부의 피쳐맵의 가중합으로 히트맵을 구성하기에 본래의 활성화 영역보다 넓은 영역이 마스킹 되는 경향이 있었다. 이 과정에서 불필요한 소리들이 함께 청취되는 문제가 있었으며, 추후 연구를 통해 분석 결과에 영향을 준 소리만을 청취할 수 있는 개선된 활성화 영역 특정 방법을 도입할 필요가 있다.

감사의 글

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2017-0-00050, 신체기능의 이상이나 저하를 극복하기 위한 휴먼 청각 및 근력 증강 원천 기술 개발)

참고문헌

[1] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. AI magazine, 38(3), 50-57.

[2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[3] Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

[4] Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806.

[5] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Advances in Neural Information Processing Systems (pp. 9505-9515).

[6] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[7] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene cnns. arXiv preprint arXiv:1412.6856.

[8] Wu, Y., & Lee, T. (2019, May). Enhancing sound texture in CNN-based acoustic scene classification. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 815-819). IEEE.

[9] McDonnell, M. D., & Gao, W. (2020, May). Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 141-145). IEEE.