음색 러닝을 위한 합성 곱 신경망 모델 분석

박소현 [1], 임선영 [1], 박영호 [1,*]
[1] 숙명여자대학교 IT 공학과
e-mail : {shpark, sunnyihm, yhpark}@sookmyung.ac.kr
*교신저자

A Study on Sound Timbre Learning Using Convolutional Network

So-Hyun Park[1], Sun-Young Ihm[1], Young-Ho Park[1,*]
[1]Dept. of IT Engineering, Sookmyng Women's University

요 약

서로 다른 음성 데이터 분류를 위한 연구는 많이 진행되고 있지만 개인이 갖고 있는 목소리 또는 각 악기들이 갖고 있는 음색 러닝 연구는 부족한 실정이다. 본 논문에서는 음색 러닝을 위한 합성 곱 신경망 분석 연구를 진행한다. 음색이란 음정과 세기가 같을 경우에도 두 소리를 구분할 수 있는 복합적인 요소이다.

1. 서론

최근 빅데이터 등장과 함께 딥 러닝을 이용한 데이터 분석이 각광받으면서, 다양한 분야에 딥 러닝을 활용하는 연구들이 등장하고 있다. 다양한 분야 중 음성 데이터 처리를 위한 합성 곱 신경망 연구를 진행한다.

서로 다른 음성 데이터 분류를 위한 연구는 많이 진행되고 있지만 개인이 갖고 있는 목소리 또는 각 악기들이 갖고 있는 음색 러닝 연구는 부족한 실정이다. 음색이란 음정과 세기가 같을 경우에도 두 소리를 구분할 수 있는 복합적인 요소들을 의미한다[1]. 본 논문에서는 음색 러닝을 위한 합성 곱 신경망을 분석한다. 음색 러닝은 음악 정보 검색 분야, 악기 인식 분야, 스피치 인식 분야에서 널리 사용되고 있다[2-5].

본 논문의 구성은 다음과 같다. 2 장에서는 기존의 다양한 합성 곱 신경망 모델을 분석한다. 3 장에서는 데이터 셋 수집방안에 대해 소개하고 마지막으로 4 장에서는 결론 및 향후 연구에 대해 소개한다.

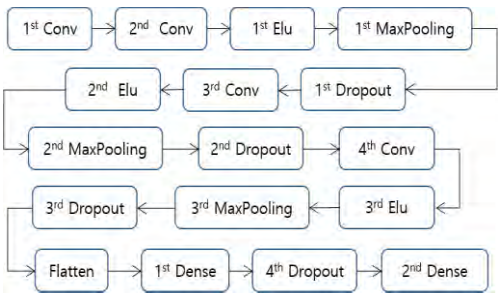
2. 음색 러닝을 위한 합성곱 신경망 모델 제안

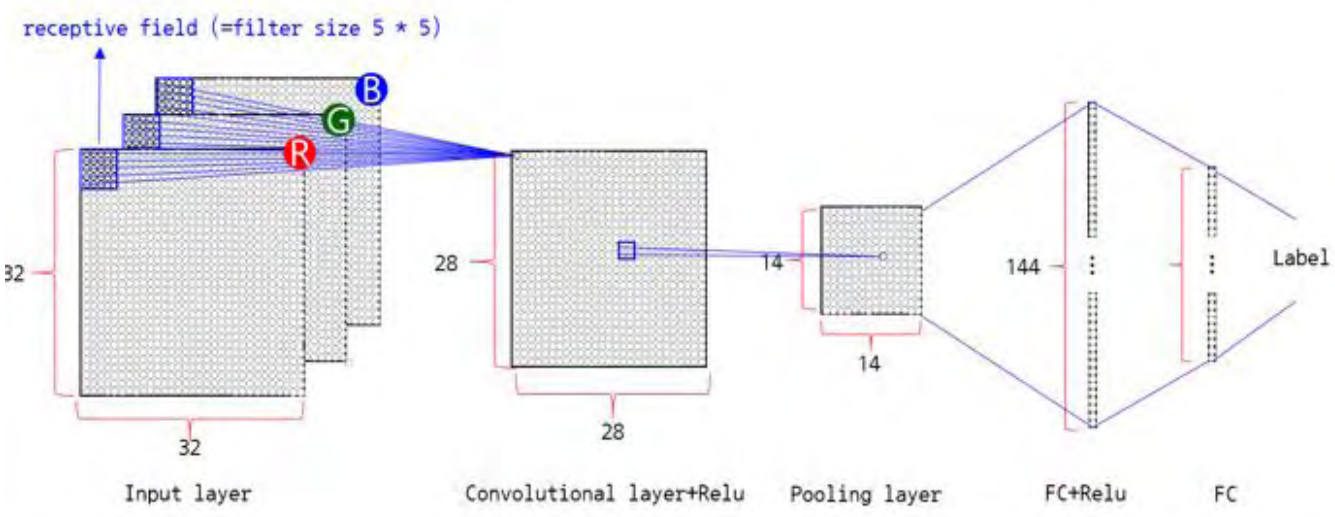
본 절에서는 합성 곱 신경망을 분석한다. Conv 는 합성 곱 레이어를 의미한다. MaxPooling 은 과 적합 문제 및 컴퓨터 자원 사용을 줄이기 위한 방법으로써 인접한 픽셀 중 가장 큰 값을 픽셀들의 대표 값으로 사용하는 Max 방법이나 가장 작은 값을 픽셀들의 대표 값으로 사용하는 Min 방법 등이 있다. Dropout 은 과 적합 문제나 컴퓨터 자원 사용 등을 줄이기 위한 방법으로써 레이어 들간 일부 연결을 강제로 끊는 방법이다.

(그림 1)의 합성곱 신경망 모델은 총 4 개의 합성곱 레이어를 사용하였고, 총 4 개의 활성화 함수로 Elu 를 사용하였다. 또한, 총 3 개의 MaxPooling 함수를 사용하였고 총 4 개의 Dropout 함수를 이용하였다. 후반 부에 Fatten 함수를 사용하고 두 번의 Dense 함수를 이용하였다.

(그림 2)는 또다른 합성 곱 신경망 예시이다. Receptive field 는 필터의 사이즈를 의미하며, 실제 데이터를 담고 있는 필터 하고는 다른 개념이다. 컬러 이미지의 경우 RGB 데이터에서 연산을 더한 값이 다음 합성 곱 레이어의 뉴런에 반영된다. 두 번째 레이어인 합성 곱 레이어와 활성화 레이어다. 활성화 레이어에서는 해당 뉴런의 값에 따라서 활성화를 시킬 것인지에 대한 여부를 결정한다. 다음 레이어는 처리해야할 데이터의 사이즈를 줄이고 오버 피팅을 피하기 위한 Pooling 레이어이다. 그 다음 레이어는 추출된 특징 벡터를 1 차원 배열로 변경하는 Fully Conneted 레이어이다. 해당 레이어에서도 활성화 함수를 사용할 수 있고, 합성 곱 레이어나 Pooling 레이어처럼 여러 겹으로 쌓을 수 있다. Fully Conneted 레이어에 저장된 Vector 값에 따라 어떤 Label 에 해당될지가 결정된다.

(그림 1) 합성곱 신경망 예시 1





(그림 2) 합성곱 신경망 예시 2

3. 데이터 셋 수집 방안

정확한 음색 러닝을 위해서는 같은 조건의 데이터를 수집하여야 한다. 따라서, 곡 하나를 여러 명이 연주하는 데이터를 수집하였다. 피아니스트는 그림 3와 같이 Yuja Wang, Christian Zimmermann, Evgeny Kissin, Natalie Schwamova 이다.

(그림 3) 네 명의 클래식 피아니스트 [6-9]

4. 결론

본 논문에서는 음색 러닝에 자주 활용되는 합성곱 신경망 모델을 분석하였다. 이를 위해 합성 곱 신경망 예시 1,2 를 들고 각 구조를 분석하였다. 분석 결과 이미지를 잘 분류하는 합성 곱 모델은 이미지화 시킬 수 있는 오디오를 분류하는데도 적합한 것으로 사료된다. 또한, 향후 연구에서는 수집한 데이터 셋을 실제로 합성곱 신경망에 적용해보고 음색 러닝 가능성을 확인해보고자 한다.

사사문구

본 연구는 산업통상자원부와 한국산업기술진흥원의 “국제공동기술개발사업”의 지원을 받아 수행된 연구결과임.

참고문헌

[1] Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. The Journal of the Acoustical Society of America, 114(5), 2946-2957.
[2] Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017, August). Timbre analysis of music audio signals with convolutional neural networks. In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 2744-2748). IEEE.
[3] Hamel, P., & Eck, D. (2010, August). Learning features from music audio with deep belief networks. In ISMIR (Vol. 10, pp. 339-344).
[4] Tardieu, D., & Rodet, X. (2007, October). An instrument timbre model for computer aided orchestration. In 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (pp. 347-350). IEEE.
[5] Cosi, P., De Poli, G., & Lauzzana, G. (1994). Auditory modelling and self- organizing neural networks for timbre classification. Journal of New Music Research, 23(1), 71-98.
[6] Yuja-Wang 연주 영상, https://www.youtube.com/watch?v=4l1bs5hlnYk
[7] Christian Zimmermann 연주 영상, https://www.youtube.com/watch?v=Ce8p0VcTbuA
[8] Evgeny Kissin 연주 영상, https://www.youtube.com/watch?v=yt2Dmg4ebh8
[9] https://www.youtube.com/watch?v=2uvAewYkEFU