음정 및 악기 분류를 위한 신경망

한철호O, 이성태, Heidi L Tessmer, 장병탁
서울대학교 컴퓨터공학부
{chhan, stlee, htessmer, btzhang}@bi.snu.ac.kr

Neural Networks for Pitch and Instrument Classification

Cheolho HanO, Sungtae Lee, Heidi L Tessmer, Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

이 논문은 음악에서 연주되는 음정 및 악기를 분류하기 위해 신경망을 사용하였다. 먼저 MIDI 파일을 임의로 생성하고 이를 오디오 파일과 피아노 롤(piano roll)로 변환하였다. 다층 퍼셉트론(multilayer perceptron), 합성곱 신경망(convolutional neural network)을 활용하여 생성된 오디오의 스펙트로그램(spectrogram) 상의 매 시점에 대해 음정 및 악기를 분류하였다. 음정 및 악기를 분류하기 위한 다양한 신경망 중에 합성곱 신경망을 사용함으로써 높은 정확도를 얻을 수 있었다.

1. 서 론

시공간적 데이터를 분석하는 문제는 인공지능에서 오랜 기간 다뤄져 왔다[1]. Bengio Group에 의해 제안된 딥러닝 이전에는 GMM, HMM과 같은 모델을 활용한 문제해결이 이뤄졌다. 그러던 중 딥러닝 (Deep learning)의 발전, 그 중에서도 합성곱 신경망 (Convolutional neural networks)의 성능이 입증[2]되면서 컴퓨터 비전 (Computer vision) 등 인공지능 전반에 발전이 있어왔다. 특히 딥러닝이 특징 추출에 효과적인 모델로 제시되면서 순차적 데이터 분야는 순환 신경망을 이용한 작곡 모델[3], 공간적 데이터 분야는 음악과 함께 예술의 한 분야를 이루는 미술 영역에서는 화가의 화풍을 모사한 모델[4]이 고안되었다. 하지만 DNA와 같은 생물학 데이터를 활용해 연구하는 생물정보학이나, 대화 시스템 설계 같은 순차적 데이터를 다루는 문제들은 여전히 어려운 문제로 남아있다. 음악은 그 중에서도 복잡한 차원을 가진 데이터로서 음악 정보 검색 (Musical Information Retrieval) 연구가 활발히 진행되어 있다. 이 분야에서는 음정[5] 및 악기[1] 분류 문제, 음악 생성[6] 등의 문제를 다룬다.

앞서 음정과 악기를 분류하고자 했던 연구는 크게 딥러닝 이전의 접근과 딥러닝 이후의 접근으로 나눌 수 있다. 딥러닝 이전의 접근에는 가우스 혼합 모델 (Gaussian Mixture Model)과 서포트 벡터 머신 (Support Vector Machine)을 이용한 접근[1]이 있고, 딥러닝 이후에는 합성곱 신경망을 적용하거나[7] 아직 음악에는 적용되지 않았지만 음성 인식 분야에서 순환 신경망을 이용해 시간 특징을 추출하는 시도가 이뤄졌다[8]. 하지만 그 중 이미지에서의 합성곱 신경망만큼 성공적이고 보편적인 모델은 없었다. 대부분 해당 문제에 대한 Supervision을 이용해 모델을 만들거나 정확도가 높지 않은 경우였다.

본 연구에서는 다층 퍼셉트론과 합성곱 신경망을 통해 스펙트로그램 상의 매 시점에 대해 음정과 악기를 나타내는 특징을 추출해 음정 및 악기 분류기를 만들어 보았다. 이를 통해 단순한 분류기를 만드는 것을 너머 음악의 일반적인 특징을 추출하고자 하였다.

2. 데이터 및 알고리즘

2.1 데이터

신경망 모델 사용을 위한 데이터는 직접 생성하여 사용하였다. 먼저 A4 (440 Hz) - A6 (1960 Hz) 의 25개의 음과 무음을 갖는 4분 길이의 단선율의 (monophonic) MIDI 파일을 임의로 생성하였다. 이를 시간에 따른 음정의 변화를 나타내는 피아노 롤 (piano roll)로 변환하고 신시사이저 프로그램인 FluidSynth[9]를 통해 샘플링 (sampling) 주파수 $Fs = 22{,}050$ Hz로 1 채널 오디오 파일 (WAV)로 변환하였다. 오디오 파일은 다시 단시간 푸리에 변환 (Short-time Fourier transform, STFT)을 통해 신호의 시간 별 주파수 성분을 나타내는 스펙트로그램 (spectrogram)으로 변환하였다. 이때 STFT 창문 (window) 길이는 $N = 2^{12}$, 도약 (hopping) 길이는 $H = N/4 = 2^{10}$으로 하였다. 에일리어싱(aliasing)과 스펙트럼 누출 (spectral leakage) 등을 막기 위해 샘플링 된 신호인 샘플들 (sample)은 창문 함수와 곱해지는데, 이때 길이가 N인 비대칭 한 창문 (asymmetric Hann window)을 사용하였다. 이렇게 되면 곡의 길이가 4분일 경우, 샘플의 수는

$$M = 4 \times 60 \times Fs = 5{,}292{,}000$$

이 된다. 또한 스펙트로그램의 크기는

$$(N/2+1) \times (\lceil M/H \rceil) = (2^{11}+1) \times 5168$$

이 된다. 그리고 주파수 해상도는 $Fs/N \approx 5$ Hz, 시간 해상도는 $H/Fs = 0.0464$ sec이 된다. 전체 샘플스펙트로그램은 역 단시간 푸리에 변환 (Inverse short-time Fourier transform, ISTFT)을 통해 다시 오디오 파일로 거의 원상태로 변환될 수 있다. 이때 만들어진 스펙트로그램이 신경망 모델의 입력으로 사용되었고, 음정 (피아노 롤로 나타냄)과 악기 (MIDI 파일의 변수로 설정되어 실제로 오디오 파일에 적용됨)가 신경망 모델의 목표 출력으로 사용되었다. 각각의 악기들의 종류가 비슷할 경우 악기 분류기 훈련이 제대로 이뤄지지 않을 수 있으므로 악기로는 음정이 있는 여러 악기 군에서 골고루 13 개 (piano, dulcimer, acoustic guitar, harp, harpsichord, violin, cello, trumpet, French horn, flute, ocarina, alto saxophone,

clarinet)를 선정하였다.

음정 분류기와 악기 분류기에서는 각각 1200개와 악기별 50개의 음악 파일을 사용하였고, 하나의 음악 파일은 4분짜리로, 그 중 5000개의 프레임을 이용하였다. 만들어진 데이터를 모델에 입력하기 위해 데이터를 섞는 과정이 필요하였다. 하나의 음정은 시간상으로 이웃한 음정에도 영향을 끼치고 시간 프레임을 나누는데 있어 박자와 음정의 길이도 영향을 끼칠 수 있기 때문에 분류기가 정확하게 작동하기 위해서는 무작위로 음정의 순서와 길이, 그리고 박자를 변형하여 훈련시킬 필요가 있었다.

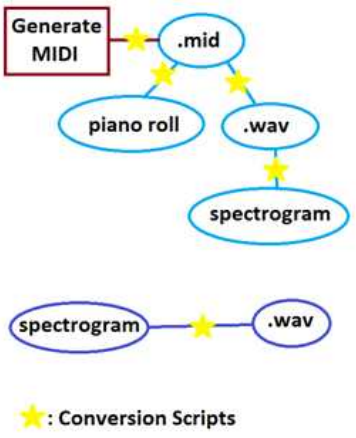

그림 1 데이터 생성 과정

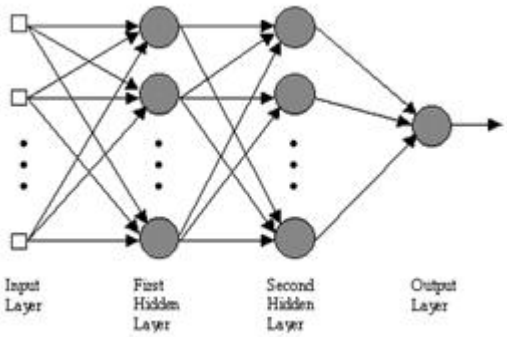

그림 2 다층 퍼셉트론 구조

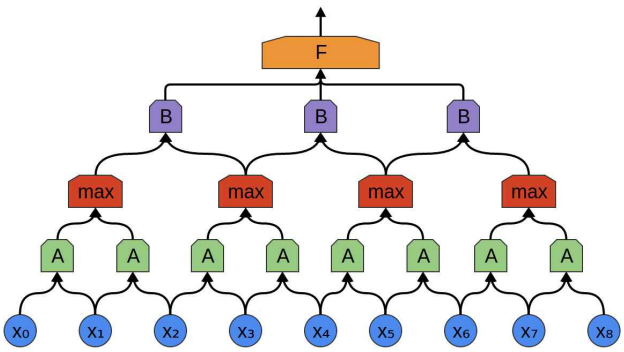

그림 3 1차원 합성곱 신경망 구조

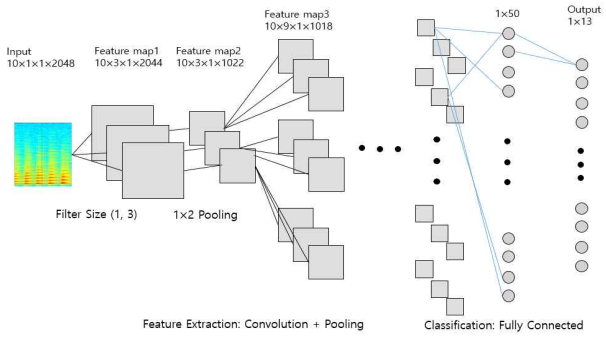

그림 4 1차원 합성곱 신경망 구조 - 악기 분류기

2.2 알고리즘

2.2.1 음정 분류기

음정 분류기는 스펙트로그램으로부터 특정 시간대에 나는 소리의 음정을 분류한다. 이때 MIDI 파일이 단선율이기 때문에, 특정 시간대에는 한 하나의 음정만 있게 되어 분류기는 다중 클래스 분류 (multiclass classification)를 수행한다. 그리고 스펙트로그램 상의 시간 간격과 음이 연주되는 시간 간격이 일치하지 않아, 스펙트로그램 기준으로 특정 시간동안 더 길게 연주되는 음을 목표 출력으로 설정하여 분류한다. 이때 분류기로는 다층 퍼셉트론 (1-8 개의 은닉층)과 7층의 1차원 합성곱 신경망을 사용해 보았다. 이미지 데이터에 주로 사용되는 필터(filter)가 2차원의 사각형 모양인 2차원 합성곱 신경망과 달리, 1차원 합성곱 신경망은 스펙트로그램의 주파수 축을 따라 1차원의 선형으로 구성되었다.

2.2.2 악기 분류기

악기 분류기는 스펙트로그램으로부터 특정 시간대의 소리를 낸 악기를 분류한다. 음정 분류기와 마찬가지로 다중 클래스 분류를 수행한다. 그러나 MIDI 파일의 특성상 악기가 한 보표 (staff) 내에서 바뀌지 않으므로, 한 곡은 하나의 악기로 연주된다. 즉, 스펙트로그램의 어떤 시간대에서도 여러 악기가 겹쳐서 나타나는 일이 없다. 따라서 음정 분류기와 달리 목표 출력이 명확하다. 다만, 스펙트로그램과 악기 쌍이 분류기의 입력과 목표 출력으로 활용될 때는 시간 단위로 나눠진 후 그 순서가 다른 스펙트로그램들과 함께 무작위로 섞이게 된다. 분류기로는 1차원 합성곱 신경망을 활용하였다.

3. 실험 및 결과

음정 분류기를 구성하기 위해 다층 퍼셉트론의 은닉층의 개수를 1-8까지 조절해보고, 직접 생성한 데이터 대신 실제 가요의 MIDI 데이터를 활용해보고 (그림 4에서 MLP-Alt), 1차원 합성곱 신경망을 적용해보는 등 다양한 실험을 하여 최대 85%의 정확도를 얻었다.

악기 분류기를 구성하기 위해 1차원 합성곱 신경망을 적용하였다. 총 13개의 악기에 대해 90%의 정확도를 얻었다.

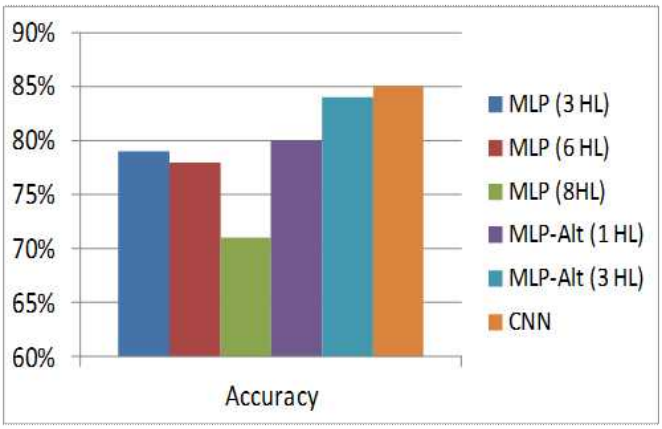

그림 5 음정 분류 정확도. n 개의 은닉층 (n HL)을 갖는 다층 퍼셉트론 (MLP), 실제 가요의 MIDI 데이터를 활용한 다층 퍼셉트론 (MLP-Alt), 1차원 합성곱 신경망 (CNN)을 적용하였다.

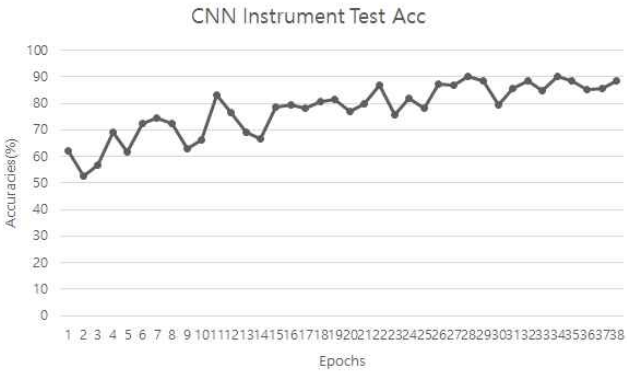

그림 6 악기 분류기 학습 곡선. 그림4에 묘사된 1차원 합성곱 신경망을 이용해 학습 횟수에 따른 분류 정확도를 그래프로 나타냈다. 최대 정확도는 90%였다.

4. 논의

본 연구에서는 음악 파일의 스펙트로그램을 분석하여 음정과 악기 분류를 수행하는 모델을 만들어 보았다. 일반적인 신경망인 다층 퍼셉트론을 먼저 이용해보았고, 이미지 데이터에 성공적으로 활용되는 합성곱 신경망을 이용해 스펙트로그램으로부터 특징 추출을 진행하였다. 실험을 통해 합성곱 신경망이 가장 좋은 정확도가 나옴을 확인할 수 있었는데, 이 때 활용한 모델은 1차원 합성곱 신경망으로 시간 특징 추출을 활용하지 않았음을 알 수 있다. 다층 퍼셉트론의 테스트 정확도가 합성곱 신경망에 비해 못 미친(~80%) 것으로 보아 합성곱 신경망이 스펙트로그램의 주파수에 대한 공간적 특징을 더 잘 추출한다는 것을 알 수 있다.

5. 결론 및 향후 연구

음정 및 악기를 분류하기 위해 다양한 신경망 중에 다층 퍼셉트론, 1차원 합성곱 신경망, 순환 신경망을 적용한 결과, 합성곱 신경망을 사용함으로써 가장 높은 정확도를 얻을 수 있었다. 향후 스펙트로그램 내의 음정 및 악기 정보의 시간에 따른 변화를 고려하여 정확도를 더 높일 수 있을 것으로 기대된다. 또한 음정, 악기뿐만 아니라 다양한 음악적 요소를 분류하기 위해 신경망을 적용해 볼 수 있다.

감사의글

이 논문은 2016 년도 정부(미래창조과학부, 국방부)의 재원으로 정보통신기술진흥센터(R0126-16-1072-SW 스타랩), 한국산업기술평가관리원(10044009-HRI.MESSI, 10060086-RISF), 국방과학연구소(UD130070ID-BMRR)의 지원을 받았음.

참고문헌

[1] Marques, Janet, and Pedro J. Moreno. "A study of musical instrument classification using gaussian mixture models and support vector machines." Cambridge Research Laboratory Technical Report Series CRL 4 (1999).

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[3] Sturm, Bob L., et al. "Music transcription modelling and composition using deep learning." arXiv preprint arXiv:1604.08723 (2016).

[4] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A neural algorithm of artistic style." arXiv preprint arXiv:1508.06576 (2015).

[5] Smaragdis, Paris, and Judith C. Brown. "Non-negative matrix factorization for polyphonic music transcription." Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.. IEEE, 2003.

[6] Boulanger-Lewandowski, Nicolas, Yoshua Bengio, and Pascal Vincent. "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription." arXiv preprint arXiv:1206.6392 (2012).

[7] Park, Taejin, and Taejin Lee. "Musical instrument sound classification with deep convolutional neural network using feature fusion approach." arXiv preprint arXiv:1512.07370 (2015).

[8] Graves, Alex, and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks." Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014.

[9] Bianchini, S.. Hanappe, P., and Lee, J. Fluidsynth: A realtime soundfont software synthesizer, 2012 URL: http://fluidsynth.elementsofsound.org/