사용자 감성 인지 기반 Latent Vector 재학습을 활용한 음악 생성 모델

장소영, 손봉기*, 이재호

덕성여자대학교, *서원대학교

thdud030101@duksung.ac.kr, *bksohn@seowon.ac.kr, izeho@duksung.ac.kr

A Music Generation Model Using Re-trained Latent Vector Conditioned on User Emotion Recognitionㄴ

Soyoung Jang, Bong-Ki Son*, Jaeho Lee

Duksung Women's Univ., *Seowon Univ.

요 약

최근 음악 생성 인공지능 모델이 많아지면서 많은 사람이 음악 관련 전공 지식이 없어도 음악 생성을 할 수 있게 되었다. 하지만, 기존 음악 생성 인공지능 모델은 사용자가 자신의 감성을 인지하고 조건으로 넣어 원하는 스타일의 음악을 생성하기 어렵다. 보통 음악 생성 인공지능 모델은 사용자가 가시적으로 확인할 수 없는 latent space에서 latent vector를 추출한 후, 무작위로 음악을 생성한다. 그렇기에 생성 모델에게 조건을 주었을 때, 기존의 음악 생성 방식이 어긋나 음악의 요소를 잃거나 조건 적용이 이루어지지 않는다. 따라서 본 논문에서는 사전 학습된 MusicVAE의 latent vector를 Actor-Critic 학습 방식으로 재학습하여 음악을 생성할 때 음악의 전체적인 구조를 따르면서 조건의 적용을 유도한다.

Ⅰ. 서 론

최근 인공지능 생성 모델이 발전하면서 음악 생성 모델도 관심을 받고 있다. 특히, 원시 오디오 파형을 사용해 음악을 생성하는 End-to-End 방식으로 뛰어난 성능을 보인다.[1] 그러나 End-to-End 방식의 음악 생성 인공지능 모델은 고차원의 데이터를 그대로 사용하고 있어서 사용자의 의도를 초기 계층부터 반영할 수 없다. 내부 메커니즘도 복잡해 모델이 특정 의도를 어떻게 유도하는지 파악에 어려움이 있다. 이런 한계점을 극복하기 위해 본 논문에서는 MusicVAE[2]를 사용해 latent vector에 조건의 적용을 유도하는 방식을 제안한다.

본 논문에서는 조건으로 사용자의 감성을 의도하도록 제안한다. 사용자의 감성을 loss function의 기준값으로 사용할 수 있도록 감성에 대한 음악의 차이를 pitch와 density로 정의하였다. 감성은 격렬함, 밝음, 발랄함, 평화로움, 청량함, 신비로움, 우울함, 맑음, 암울함으로 구분하였다. 사용한 음악 데이터를 위의 감성으로 구분하고, 각 음악의 평균 pitch와 평균 density 값을 추출하였다. 따라서 생성된 음악은 Actor-Critic 방식으로 학습이 진행되면서 latent vector에서 특정 pitch와 density가 유도되면서 해당 감성에 맞는 음악이 생성되게 된다.

Ⅱ. 본론

본 논문에서는 음악의 전체적인 구조와 연속적인 구조를 학습하기 위해 사전 학습이 완료된 MusicVAE의 latent vector를 사용하였다. MusicVAE는 Google의 magenta에서 개발한 모델로 16마디의 melody를 생성하는 'hierdec-mel_16bar'를 활용해 연구를 진행하였다.

2.1 MusicVAE

MusicVAE는 Long Short-Term Memory(LSTM) 네트워크를 기반으로 MIDI 파일을 처리하여 음악의 멜로디와 조화를 추출해 새로운 음악을 생성하는 모델이다. 음악의 장기적인 구조와 스타일을 학습하는 데에 주로 사용된다. 음악의 다차원적 속성들을 학습하여 저차원의 잠재 공간에 대응하면서 음악의 기본적인 요소들과 구조를 학습한다. 음악적 아이디어들을 새로운 방식으로 결합하고 변형하여 창의적인 작곡 프로세스를 지원하면서 기존 음악의 스타일을 모방하고 새로운 음악 조각을 생성하는 것이 가능하다. 따라서 본 논문에서는 MusicVAE를 사전 학습해서 음악의 장기적인 구조를 파악한 후, 잠재 벡터의 재학습으로 사용자의 조건을 포함해 음악을 생성할 수 있도록 하였다.

2.2 Loss Function 정의

음악의 감성 분류를 진행해 감성에 맞는 pitch와 density의 기준값을 임의로 설정하였다. 이 기준값으로 각각 따로 loss function 정의를 진행하였다. Pitch loss function은 생성된 음악의 개별 음표의 적합성을 평가하도록 구성하였다. Density loss function은 생성된 음악의 특정 시간 간격 동안의 음표 밀도를 평가하도록 구성하였다.

Pitch loss function은 생성된 음표가 화음의 기본음, 3도 화음, 5도 화음에 해당할 때, 높은 가중치를 가질 수 있도록 설정하였다. 화음에 어울리는 음이 생성되면 작은 손실값을 가지도록 하여 보상을 주고, 화음에 어울리지 않는 음이 생성되면 큰 손실값을 부여해서 페널티를 가질 수 있도록 설정하였다. 화음에 적합한 음악을 생성할 수 있도록 loss function을 구성하였다.

Density loss function은 생성된 음악의 spectrogram을 분석해 특정 시간 동안 에너지를 추출해 기준값과 비교하도록 구성하였다. 너무 많거나 적은 음표가 집중되지 않도록 설정하여 생성된 음악이 특정 밀도를 유지할 수 있도록 하는 loss function으로 구성하였다.

2.4 Latent Vector 재학습

본 논문에서는 MusicVAE의 사전 학습을 통해 음악의 전체적인 구조 및 연속적인 구조를 이해하는 latent vector를 추출하였다. 추출된 latent vector를 Actor-Critic 방식으로 재학습을 진행하면서 사용자의 의도가 latent vector에 유도되도록 진행하였다. Actor는 latent vector를 생성된 latent space 내에서 재식별을 하면서 조건을 유도하도록 진행된다. Critic은 Actor가 생성한 latent vector가 음악의 전체적인 구조 및 연속적인 구조를 따르고 있는지와 pitch와 density로 정의된 loss function을 중심으로 사용자가 지정한 조건을 만족하는지를 동시에 판단을 진행하도록 구성되었다. Actor와 Critic의 동시 학습으로 latent vector의 재학습이 진행된다. 재학습이 진행되고 난 후, MusicVAE의 decoder를 활용해 latent vector로 음악을 생성하는 과정을 진행하였다. 생성된 음악은 음악의 전체적이고 연속적인 구조를 만족하면서 사용자의 의도를 담도록 생성이 된다.

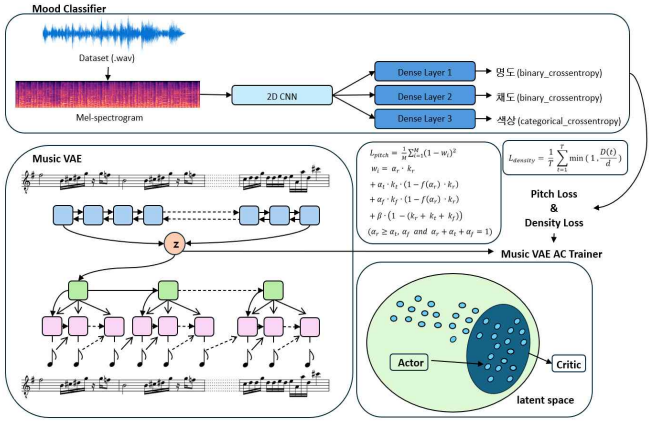

그림1. 전체 학습 과정

Ⅲ. 결론

본 논문에서는 사용자 감성 인지를 조건으로 활용해 음악을 생성하는 모델을 loss function 재정의를 통해 제안하였다. 사용자가 감성을 임의로 선택하면 해당하는 pitch와 density의 값을 추출해 loss function의 기준값으로 사용하도록 하였다. 사전 학습된 MusicVAE에서 latent vector를 추출하고, loss function을 중심으로 Actor-Critic 학습 방식으로 latent vector의 재학습을 진행하였다. 재학습된 latent vector를 MusicVAE의 decoder를 활용해 음악을 생성할 수 있도록 하였다. 향후, pitch와 density 외의 다양한 loss function 재정의를 통해 보다 더 다양하게 사용자의 의도를 표현할 수 있도록 구현한다면, 더 다양한 음악을 생성할 수 있다고 기대된다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원받아 수행된 연구임(과제번호- 2022R1A2C1009951).

참 고 문 헌

[1] Dhariwal, Prafulla, et al. "Jukebox: A generative model for music." arXiv preprint arXiv:2005.00341 (2020).

[2] Roberts, Adam, et al. "A hierarchical latent vector model for learning long-term structure in music." International conference on machine learning. PMLR, 2018.