Pitch 및 Density 손실 함수의 설계 및 최적화에 관한 연구: MusicVAE와 결합된 Actor-Critic 방식을 활용한 사용자 의도 기반 음악 생성 모델

장소영, 이재호

덕성여자대학교

thdud030101@duksung.ac.kr, izeho@duksung.ac.kr

Design and Optimization of Pitch and Density Loss Functions: A Study on User-Intent Based Music Generation Model Combining Actor-Critic Approach with MusicVAE

Soyoung Jang, Jaeho Lee

Duksung Women's Univ.

요 약

인공지능을 활용한 음악 생성 모델 중에서 사용자의 의도를 포함시켜 음악을 생성하는 것은 복잡한 과제이다. 본 논문은 사전 학습된 MusicVAE의 latent vector를 추가 학습을 진행하는 방식의 사용자 의도에 따른 음악 생성 모델을 제안한다. 사용자의 의도는 pitch와 density를 이용해 손실 함수로 표현해 사용한다. 추가 학습은 강화 학습 중 Actor-Critic 방식으로 Actor가 latent vector를 수정하고 Critic이 손실 함수로 평가하면서 사용자의 의도가 음악에 반영될 수 있도록 하는 모델을 연구하였다. 본 연구는 사용자의 의도를 가진 음악을 생성하는 인공지능 모델 분야에 기여를 할 것이다.

Ⅰ. 서 론

음악은 멜로디, 리듬, 화음 등 여러 요소로 구성된 예술이다. 모든 요소를 고려하고 적절한 화합이 이루어져야 아름다운 음악을 생성할 수 있다. 그러나 인공지능을 활용한 작곡 분야에서 음악의 요소를 고려하기엔 복잡하다. 사용자가 원하는 의도를 이해하고 인공지능을 활용한 작곡이 어려운 이유가 이 복잡성이다. 기존의 인공지능을 활용한 음악 생성 모델에서 사용자의 의도를 넣으려고 하는 많은 시도가 있었지만, 학습을 진행하면서 사용자의 의도는 음악 생성에서 고려되어야 하는 중요한 요소가 아니라고 인공지능이 판단해 무시된다.

본 연구에서는 이 문제를 극복하기 위해, 사전 학습된 MusicVAE[1]를 활용한다. 사전 학습된 MusicVAE를 사용자 의도에 맞게 추가 학습을 진행하게 되는데, 이때 사용자의 의도를 손실 함수를 통해 모델에 넣는다. 본 논문에서는 pitch와 density 관련 손실 함수를 사용해 음악이 특정 화음에 어울리고, 빠르기를 조절할 수 있도록 생성한다. 사전 학습된 MusicVAE를 활용해 데이터의 latent vector를 학습하게 되고, 강화 학습 중 Actor-Critic 학습 방법을 활용해 사용자 의도에 맞춰 latent vector를 수정한다. Actor는 latent vector를 수정하고 Critic은 손실 함수에 따라 Actor의 학습을 평가한다. 이렇게 사용자의 의도를 손실 함수로 넣어 강화학습을 진행하는 음악 생성을 하는 모델을 제안한다.[2]

Ⅱ. 본론

본 논문에서는 MIDI 데이터로 학습된 MusicVAE를 활용한다. 사용자의 의도는 손실 함수를 통해 표현하며, MusicVAE에서 생성된 latent vector를 사용자의 의도에 맞게 Actor-Critic 방식으로 추가 학습을 진행하는 모델을 제안한다.

2.1 선행 연구

MusicVAE는 음악 생성을 위한 딥러닝 기반 모델로, VAE를 기반으로 한다. 음악을 입력으로 받아 latent space를 확률 분포로 학습을 진행하고 latent vector를 샘플링하면서 인코딩을 진행한다. 이 latent vector는 음악의 중요한 특징을 가지고 있다. 인코딩된 latent vector를 입력으로 활용하는 디코더는 vector를 다시 음악으로 재생성한다. 디코더가 생성한 음악과 원본 음악 간의 차이를 손실 함수로 정의해 손실 함수가 최소화하는 방향으로 모델 학습을 진행한다.

Actor-Critic[3]은 강화 학습에서 사용되는 알고리즘 중 하나로, Actor와 Critic, 두 모델이 상호 작용하면서 최적의 행동을 학습하는 방식이다. Actor는 주어진 환경에서 어떤 행동을 선택해야 하는지 결정하는 확률 함수이다. 학습을 통해 최적의 행동을 찾아낸다. Critic은 Actor의 결정을 평가하면서 현재 성능을 알려준다. 주로 Q-learning, TD-learning 등의 알고리즘을 사용해 학습한다.

2.2 모델 제안

본 논문에서 제안한 모델은 MusicVAE를 사전 학습해서 사용한다. 사전 학습을 진행하면, 음악의 특성과 구조가 고려된 latent vector를 얻을 수 있으며, 이 latent vector는 음악의 특징을 담고 있다. latent vector를 디코더에 통과시킨다면 음악을 재생성할 수 있다. 생성된 latent vector의 특정 차원을 축소하거나 확장하면서 음악의 특성을 변화시킬 수 있다. 본 모델에서는 이 과정을 추가 학습을 통해 진행한다.

특정 차원의 축소 혹은 확장 과정에서 사용자의 의도를 label로 설정한다면, latent space 내에 사용자의 의도가 반영되어 latent vector를 특정 의도에 학습하는 과정이 수월하다. 그러나 음악의 경우, label을 설정하기에 기준을 명확하게 확정하기 어렵다. 그렇기 때문에 label 대신,

Actor-Critic을 활용해 latent vector를 찾는 과정을 진행한다. Actor-Critic 과정에서 Actor는 latent vector의 특정 차원을 축소하고 확장하면서 사용자의 의도를 반영하는 방식을 학습한다. 본 논문에서는 음악의 pitch와 density를 사용자의 의도로 정의해 손실 함수에 사용한다. Actor는 latent vector를 수정하면서 새로운 음악을 생성하는데, 기존 MusicVAE의 음악 생성의 일반성과 사용자의 의도가 담긴 다양성이 모두 고려된다. Critic의 경우, Actor가 수정한 latent vector와 손실 함수를 기반으로 Actor의 학습을 평가한다. 학습을 평가하면서, Actor가 사용자 의도를 더 반영할 수 있도록 하는 방향으로 학습을 진행한다.

본 논문에서 제안하는 모델 방식은 사용자의 의도를 손실 함수로 정의하고, MusicVAE와 Actor-Critic 학습을 결합해 사용자의 의도를 반영하는 음악을 생성할 수 있도록 한다.

2.3 손실 함수

본 논문에서는 사용하는 손실 함수는 아래 식과 같다.

$$L_{pitch} = \frac{1}{M}\sum_{i=1}^{M}(1 - w_i)^2$$

그림1. 음악의pitch손실함수

그림 1의 수식은 pitch를 기준으로 음악적 맥락에서 생성된 음이 적절한지를 판단하는 손실 함수이다. w는 특정 음이 주어진 음악의 맥락 내에서 어우러지는 정도를 나타내는 가중치이다. 음이 화음 집합에 속한 경우에는 어울리는 음이라고 생각해 가중치의 값을 키워 전체적인 손실값이 낮아질 수 있도록 한다. 화음 집합에 속하지 않은 경우는 음악적 맥락에 맞지 않는다고 판단해 가중치의 값을 줄여 전체적인 손실 값이 커질 수 있도록 손실 함수를 설정한다.

$$\begin{gathered} w_i = \alpha_r \cdot \mathbf{1}(m_i \text{ is root}) + \\ \alpha_t \cdot \mathbf{1}(m_i \text{ is third}) \cdot (1 - \lambda_r \cdot \mathbf{1}(m_i \text{ is root})) + \\ \alpha_f \cdot \mathbf{1}(m_i \text{ is fifth}) \cdot (1 - \lambda_r \cdot \mathbf{1}(m_i \text{ is root})) + \\ \beta \cdot (1 - (\mathbf{1}(m_i \text{ is root}) + \mathbf{1}(m_i \text{ is third}) + \mathbf{1}(m_i \text{ is fifth}))) \\ (\alpha_r \geq \alpha_t, \alpha_f \\ \text{and } \alpha_r + \alpha_t + \alpha_f = 1) \end{gathered}$$

그림2. 음악의pitch손실함수의가중치계산식

그림 2의 수식은 화음 집합에 속한 경우와 속하지 않은 경우에서 가중치를 설정하기 위한 계산식이다. 해당 음의 각 기본음, 3도 화음, 5도 화음에 해당하는지를 판별해 가중치를 다르게 부여한다. 만약 해당 음이 기본음인 동시에 3도 화음 혹은 5도 화음이라면, 가중치가 높아지는 것을 방지하기 위해 감소 효과도 계산식에 적용한다. 모델이 음악적 맥락에서 더 적절한 예측을 하기 위한 pitch 관련 손실 함수와 그 가중치의 정의이다.

$$L_{density} = \frac{1}{T}\sum_{t=1}^{T}\min\left(1, \frac{\mathrm{D(t)}}{d}\right)$$

그림3. 음악의density손실함수

그림 3의 수식은 density를 기준으로 음악적 맥락에서 생성된 음의 빠르기가 적절한지를 판단하는 손실 함수이다. 시간 간격 내에서 스펙트로그램의 에너지 밀도를 측정해 사용한다. 시간 간격 동안의 에너지 분포가 일정 임계값 이하로 유지되는지를 평가한다.

Ⅲ. 결론

본 논문에서는 MusicVAE를 사전 학습하고 Actor-Critic 방식으로 추가 학습을 통해 사용자의 의도를 손실 함수에 넣어 latent vector를 수정하면서 음악을 생성하는 모델을 제안한다. 본 모델은 사용자의 의도를 손실 함수에서 사용하고, 이를 latent vector에 적용하기 위해 강화 학습을 사용하는 새로운 접근 방식을 제안한다. 현재 본 모델을 개발하고 있지만, 검증 과정을 거치면서 많은 실험이 진행되어야 한다. 향후 모델이 완성된다면, 사용자의 의도를 고려한 인공지능을 활용한 작곡 분야 발전에 기여가 예상된다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018, July). A hierarchical latent vector model for learning long-term structure in music. In International conference on machine learning (pp. 4364-4373). PMLR.

[2] Engel, J., Hoffman, M., & Roberts, A. (2017). Latent constraints: Learning to generate conditionally from unconditional generative models. arXiv preprint arXiv:1711.05772.

[3] Konda, V., & Tsitsiklis, J. (1999). Actor-critic algorithms. Advances in neural information processing systems, 12.