음악 생성 모델에 대한 연구 방법 및 방향 사례 분석

이현주, 손봉기*, 이재호
덕성여자대학교, *서원대학교
20221004@duksung.ac.kr, *bksohn@seowon.ac.kr, izeho@duksung.ac.kr

A Case Study of Research Method and Approach on Music Generation Model

Lee Hyeon Joo Bong-Ki Son*, Jaeho Lee
Duksung Women's Univ., *Seowon Univ.

요 약

생성 모델은 지난 몇 년 동안 크게 성장해왔다. 생성 모델의 목표는 데이터의 특징을 찾아내서 학습한 뒤 새로운 데이터를 생성하는 것이다. 이러한 생성 모델은 음악 산업에도 사용된다. 생성 모델 초기에는 마르코프 체인 알고리즘이 사용되었고 이후, LSTM, GAN, 트랜스포머, 디퓨전이 등장했다. LSTM은 시간적 의존성 학습에 강하지만 긴 시퀀스 처리에 한계가 있다. GAN은 생성자와 판별자의 경쟁적 학습 구조를 통해 다양한 음악 스타일을 생성한다. 트랜스포머는 셀프 어텐션 기법을 통해 시퀀스를 병렬로 학습할 수 있으며 디퓨전은 점진적으로 데이터를 변형하고 복원하는 과정을 통해 복잡한 패턴을 학습한다. 본 논문은 이런 생성형 모델 기술과 대표 모델을 소개한다.

Ⅰ. 서 론

최근 인공지능 분야가 빠르게 발전하면서 점차 다양한 산업 분야와 접목하고 있다. 음악 산업 역시 인공지능을 통해 크게 변화하고 있다. 현재 음악 산업에서 가장 활발히 활용되는 인공지능은 대표적으로 음악 작곡과 음악 추천이다.

음악은 기본적으로 2가지 특징을 지니고 있다. 연주곡인 경우는 소리의 높낮이와 빠르기와 같은 피치와 리듬의 반복인 멜로디가 있고 가사가 있는 곡인 경우는 멜로디에 추가적으로 텍스트의 특징을 지니고 있다. 이런 음악 특징을 기반으로 현재 다양한 음악 생성 모델이 나오고 있다.

음악 생성 모델 계보의 초기는 MIDI (Musical Instrument Digital Interface) 기반으로 사용하는 모델이 존재했다. 그 다음 시퀀스 기반인 마르코프 체인 (Markov Chain) 알고리즘을 이용하여 멜로디를 생성했다. 2000년 중반부터는 신경망을 활용한 LSTM (Long Short-Term Memory)와 같은 순한 신경망 (RNN) 을 이용했다. 그리고 2010년대 초반과 중반에는 GAN (Generative Adversarial Network) 과 VAE (Variational Autoencoder)를 사용하여 더 효율적으로 모델을 학습했다. 그 이후에는 Transform과 Diffusion 기반으로 모델을 생성했다.

본 논문에서 위에 소개한 기술 중 인공신경망 기반인 대표적인 모델과 음악 생성 모델의 기술 방향성을 다룬다.

Ⅱ. 음악 생성 기술과 대표적인 음악 생성 모델

2.1 마르코프 체인 알고리즘

마르코프 체인(Markov Chain)은 마르코프 성질을 가진 이산 확률이다. 이는 특정 상태의 확률은 오직 과거 의존한다는 성질을 이용한다. 마르코프 체인은 음악 생성에 크게 활용할 수 있다. 멜로디는 앞의 노트와 뒤의 노트 간의 상관관계가 존재한다. 따라서 마르코프 체인 알고리즘 기반으로 다음 멜로디의 확률을 구해서 모델링할 수 있다. 하지만 마르코프 체인 알고리즘은 전체 흐름을 무시하고 오직 현재 상태만을 고려하기 때문에 한계가 존재한다.

2.2 LSTM

음악은 연속된 음표와 가사로 이루어져 있어 이와 같은 연속적인 특징을 잘 활용할 수 있는 순환 신경망 (RNN, Recurrent Neural Network)을 이용한 음악 생성 모델이 있다. 순환 신경망은 가장 기본적인 시퀀스 (sequence) 모델이다.

순환 신경망의 가장 큰 특징은 바로 순차적인 데이터를 연결해 정보를 파악하고 학습한다는 것이다. 즉 이전 단계의 출력 값과 현재 입력 값 두 개를 입력값으로 받는다. 하지만 이는 해당 데이터의 거리가 멀어질수록 데이터를 연결하는 기능이 약해진다. 따라서 음악의 길이가 길수록 학습 능력이 떨어진다. 이를 해결하기 위해 LSTM(Long Short - Term Memory)가 나왔다. 이는 학습할 때 중요한 정보를 골라서 저장하고 보존해서 음악 생성과 같이 오랜 시간 정보를 필요 하는 데 사용된다.

대표적으로 구글의 마젠타 (Magenta) 프로젝트에서 진행한 MusicRNN이 있다. 이는 시퀀스의 학습을 통해 새로운 멜로디와 리듬을 생성한다.[3]

2.3 GAN

음악은 여러 악기의 조합으로 멜로디를 생성한다. 이와 같은 여러 악기와 멜로디가 이루어진 음악을 만들기 위해 GAN (Generative Adversarial Network) 기술을 이용한다. GAN은 생성자 (generator)와 판별자 (discriminator) 으로 나뉘어진다. 생성자는 정답과 학습 결과를 구분하기 어렵게 데이터를 계속 생성하고 판별자는 이렇게 생성된 데이터와 정답

데이터를 구분한다. 이와 같은 경쟁 구조 시스템을 통해 생성자는 점차 정답에 가깝게 학습한다.

GAN을 이용한 대표적인 음악 생성 모델은 MuseGAN이 있다. 이 모델은 다양한 음악 스타일의 특징을 학습하여, 특정 장르에 맞는 음악을 생성한다. 또한 사용자가 특정 스타일을 지정할 수 있어 다양하고 사용자 맞춤형 음악을 생성한다.

MuseGAN은 세 가지 주요 멀티 트랙 모델 (multitrack model)과 시간 의존적 모델 (temporal model)을 사용하여 다양한 음악 요소들을 생성한다. [4]

멀티 트랙 모델은 잼잉 모델, 작곡가 모델, 하이브리드 모델로 구성되어 있다. 잼잉 모델은 여러 생성자로 구성되어 있다. 생성자들은 각자 독립적인 트랙 음악을 생성한다. 작곡가 모델은 단일 생성자로 구성되어 있으며 모든 트랙을 생성하여 음악이 전체적으로 조화롭게 만들어준다. 마지막으로 하이브리드 모델은 잼잉 모델과 작곡가 모델을 합친 모델로 여러 생성자가 트랙 간의 조화를 이루며 각각 독립적인 음악을 생성한다. [4]

시간 구조 모델은 음악이 시간에 따라 어떻게 변화하는지 모델링하는 걸 목표로 한다. 2가지 방법으로 구현이 되는데, 첫 번째 방법은 바로 처음부터 생성 (Generation from Scrath) 방법이다. 이는 고정된 음악 길이의 구절을 생성하는 걸 목표로 한다. 이는 음악에서 정해진 박자에 따른 구간 단위로 음악을 차례대로 생성하여 전체 음악의 구절을 완성한다. 다음은 트랙 조건부 생성 (Track-conditional Generation) 방법으로 특정 트랙의 음악 구간 시퀀스가 주어졌다고 가정하고 그 트랙의 시간 구조를 학습하여 나머지 트랙을 생성하여 전체 곡을 완성한다.[4] 따라서 특정 트랙을 기반으로 나머지 트랙을 생성하여 전체적으로 조화로운 음악을 완성한다. MuseGAN은 이러한 두 가지 모델을 통합하여 다양한 악기와 스타일의 음악을 한 번에 생성한다.

2.4 Transformer

반복은 문학, 연극, 미술, 음악 등 다양한 장르에서 사용된다. 음악에서는 반복은 중요한 역할을 한다. 이는 청자에게 음악의 인식, 이해, 감정 전달과 기억에 많은 영향을 기여한다.

트랜스포머에서 사용하는 셀프 어텐션 기술은 이러한 반복 구조를 잘 발견한다. 따라서 트랜스포머를 이용하여 반복되는 음악을 생성할 수 있다.[2]

트랜스포머는 주로 자연어 처리에서 중요한 역할을 한다. 대부분 자연어 언어 처리는 인코더와 디코더 구조를 가진 순환 모델인데 이는 시퀀스가 길수록 메모리 문제로 인해 병렬적 처리가 어려운 한계를 가지고 있다. 하지만 트랜스포머는 순환 없이 어텐션 기법 기반을 이용해서 의존성을 찾는다.[1] 이는 재귀적으로 모든 시퀀스를 각자 처리하지 않고 오직 행렬곱을 이용해서 병렬적으로 시퀀스 데이터를 처리해서 처리 속도가 빠르다. 그래서 장기간 의존성을 잘 학습하고 시퀀스 변환 문제도 해결한다. 이러한 트랜스포머 기술은 음악 생성에도 효과적으로 활용된다. 이는 음악의 시퀀스 데이터를 처리할 때 고유한 시간 관계와 패턴을 효과적으로 학습할 수 있다. 대표적인 트랜스포머 모델은 OpenAI의 Jukebox와 구글 마젠타 팀에서 개발한 Music Transformer가 있다.

Jukebox는 트랜스포머 기반 모델로 다양한 장르와 스타일에서 고품질 음악을 생성한다. 이는 오디오 데이터 자체를 직접 처리하며, 노래의 멜로디, 가사, 보컬 스타일등 여러 음악적 요소를 통합하여 새로운 곡을 생성한다. VQ-VAX-2 (Vector Quantized Variational AutoEncoder) 구조를 사용하여 오디오 데이터를 처리한다. [5]

Music Transformer는 음악의 재현성과 멜로디 생성에 초점을 맞춘 모델이다. 이는 특히 장기 음악 구조를 유지하면서 새로운 음악을 생성한다. Music transformer는 self-attention 기법을 사용한다. Self-attetion은 입력 시퀀스의 모든 위치에서 정보를 집계해서 곡 전체에서 음악 관계를 파악 하는데 도움을 준다. 그리고 기존 transformer와 다르게 절대 위치 인코딩이 아닌 상대 위치 인코딩을 도입하여 입력 시퀀스 간의 상대적 거리를 고려하여 장기 의존성에 효과적이다.[2] 따라서 장기 음악 구조를 생성하고 이는 반복적인 리듬을 인공지능이 학습하고 재현할 수 있다.

2.4 Diffusion

디퓨전은 물리학적 이론인 열역학에서 영감을 받아 데이터 분포의 변화를 학습하여 새로운 데이터를 생성하는 모델이다. 이는 마르코프 체인 기법을 사용하여 원본 데이터 분포를 파괴하는 전방 과정 (Forward Process)과 그 데이터 분포를 복원하는 역방 과정 (Reverse Process) 으로 나뉜다. [6]

전방 과정은 원본 데이터에 점진적으로 가우시안 노이즈가 추가되어 복잡한 분포로 변형시킨다. 이는 시간이 지날수록 데이터에 점차 더 많은 노이즈가 추가되어 최종적으로 순수한 노이즈만 있는 형태로 변한다. 역방 과정은 전방 과정에서 추가된 노이즈를 점진적으로 제거하여 원래 데이터로 복원하는 것이다. 이러한 전방 과정과 역방 과정 간의 분포 차이를 최소화하게 학습한다.

음악 데이터는 시간적, 주파수적 특성이 복잡하게 구성된 고차원 데이터이다. 디퓨전은 이러한 고차원 데이터 분포를 효과적으로 학습할 수 있다. 먼저 점진적인 변화 학습을 통해 음악과 같이 복잡한 데이터 구조를 자세하게 모델링하고 멜로디 간의 시간적 연속성과 구조를 잘 잡을 수 있다. 그리고 디퓨전은 노이즈를 추가하고 제거하는 과정을 통해 학습하기 때문에 복잡한 분포를 학습한다. 따라서 음악의 복잡한 패턴을 잘 찾아내고 다양한 음악 스타일을 만들 수 있다. 대표적인 디퓨전 모델은 DiffWave와 WaveGrad가 있다.

DiffWave는 디퓨전 과정을 사용하여 멜-스펙트로그램을 입력으로 받아 원래 오디오 신호를 생성하는 모델이다. 이는 U-Net 구조로 되어 있어 멀티 스케일 데이터 특징을 학습하고 효과적으로 노이즈를 제거한다. WaveGrad는 파형을 생성하는 모델로 U-Net 구조가 아닌 forward 방식을 사용하여 노이즈를 제거한다.[7] WaveGrad는 DiffWave보다 구조가 간단하여 학습 속도가 빠르다. 두 모델은 모두 디퓨전 기반으로 생성된 모델로 고차원 데이터의 복잡한 분포를 학습하고 이를 통해 다양한 음악 스타일을 생성한다.

Ⅲ. 결론

본 논문은 음악 생성 모델의 기술을 설명하고 대표적인 모델을 살펴보았다. LSTM은 순환 신경망 일종으로 시간에 따라 변화하는 음악 데이터를 처리하는데 좋지만 장기의존성에 한계가 존재한다. GAN은 생성자와 판별자라는 경쟁적인 학습을 통해 다양한 음악 스타일을 생성하고 경쟁 학습 구조를 통해 진짜 같은 음악을 생성한다. Transformer는 셀프 어텐션 기법으로 시퀀스 데이터를 병렬적으로 처리한다. 따라서 장기 의존성을 학습할 수 있다. Diffusion은 데이터를 점진적으로 변형하고 이를 다시

복원하는 구조로 점진적으로 데이터를 생성하여 복잡한 구조와 패턴을 잘 학습한다.

음악 생성 모델 기술은 지금도 활발히 연구되고 있으며 인공지능을 활용한 음악 생성은 앞으로 더 다양하고 창의적인 음악을 생성하여 다양한 분야에 사용될 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Vaswani, Ashish et al. “Attention is All you Need.” Neural Information Processing Systems (2017).

[2] Huang, Cheng-Zhi Anna et al. “Music Transformer: Generating Music with Long-Term Structure.” International Conference on Learning Representations (2018). 7

[3] Roberts, Adam, et al. "MusicRNN: Generative Recurrent Neural Network for Music." Magenta. [Online]. Available: https://magenta.github.io/magenta-js/music/demos/music_rnn.html

[4] Dong, Hao-Wen, et al. "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment." MuseGAN. [Online]. Available: https://hermandong.com/musegan/model

[5] Dhariwal, Prafulla et al. “Jukebox: A Generative Model for Music.” ArXiv abs/2005.00341 (2020): n. pag.

[6] Zhu, Peng Fei et al. “ERNIE-Music: Text-to-Waveform Music Generation with Diffusion Models.” ArXiv abs/2302.04456 (2023): n. pag.

[7] Chen, Nanxin et al. “WaveGrad: Estimating Gradients for Waveform Generation.” ArXiv abs/2009.00713 (2020): n. pag.