creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사학위논문

음악적 특징 레이블 기반 심볼릭 음악 생성 네트워크

FLAGNet: Feature Label based Automatic Generation Network for Symbolic Music

2023년 2월

서울대학교 융합과학기술대학원

지능정보융합학과

고 성 현

음악적 특징 레이블 기반 심볼릭 음악 생성 네트워크

FLAGNet: Feature Label based Automatic Generation Network for Symbolic Music

지도교수 이 교 구

이 논문을 공학석사 학위논문으로 제출함

2023년 2월

서울대학교 융합과학기술대학원

지능정보융합학과

고 성 현

고성현의 공학석사 학위 논문을 인준함

2023년 2월

위 원 장: 이 원 종

부위원장: 이 교 구

위 원: 서 봉 원

요 약

자동 음악 생성 기술은 최근 몇 년 동안 매우 활발하게 연구되어 왔다. 그러나 이러한 연구에서 음악을 데이터로만 분석하는 것이 일반적이었고, 음악의 기반지식을 다루는 것은 생략되거나 어려운 작업으로 간주되었다. 특히 음악의 각 마디의 특성을 분석하고 이에 기반지식을 적용하는 과정은 사람의 작곡에서는 필수적임에도 불구하고, 자세하게 다루어지는 연구가 많지 않다. 우리는 마디의 음악적 특성과 선제 되는 음표 조건을 고려하여 각 마디를 생성함으로써 음악을 생성하는 모델을 제안한다.

먼저 심볼릭 음악 데이터를 상대적 음고 피아노 롤 표현(Relational Pitch Pianoroll Representation)으로 분석하여 피아노롤(Pianoroll) 기반 미디(MIDI) 인코딩 방법의 활용도를 높이고 생성된 결과를 음악적으로 광범위하게 사용할 수 있도록 하였으며, 이를 응용하여 다양한 이미지 기반 모델을 학습시켜 유의미한 결과를 얻어낼 수 있도록 하였다.

또한 심볼릭 음악 데이터 생성을 위해 다중 벡터 조건부 딥 컨볼루셔널 적대적 생성망(Multi-vector Conditional Deep Convolutional Generative Adversarial Network)를 사용하여 선제 되는 음표 조건과 음악적 특성 레이블을 반영하여 새로운 미디 데이터를 생성하도록 모델을 훈련했다. 또한 음악적 기술 레이블의 조합을 얻어내기 위하여 Long Short-Term Memory와 Gated Recurrent Unit을 활용하였다. 결과적으로 모델 FLAGNet은 다양한 음악적 요소를 고려하여 인상적인 심볼릭 음악을 생성할 수 있음을 보였다.

주요어: 음악 생성 모델, 심볼릭 음악, 상대적 음고 피아노 롤 표현, 다중 벡터 조건부 딥 컨볼루션 적대적 생성 네트워크
학번: 2021-24957

차 례

요 약 **i**

제 1 장 서 론 **6**

1.1 연구 배경 6

1.2 연구 목표 9

제 2 장 배경 이론 및 관련 연구 **12**

2.1 배경 이론 12

2.1.1 심볼릭 음악의 표현 방식 12

2.1.2 조건부 적대적 생성 네트워크 13

2.2 관련 연구 16

2.2.1 자동 음악 작곡 16

2.2.2 피아노 롤의 새로운 표현 방식 17

제 3 장 제안 기법 **19**

3.1 MIDI의 상대적 음고 피아노 롤 표현 19

3.2 다중 벡터 조건부 딥 콘볼루션 적대적 생성 망 21

3.2.1 모델 입력 및 출력 22

3.3 MIDI 후처리 22

3.4 FLAGNet의 전체 구성 25

제 4 장 실험 **27**

4.1 데이터셋 27

4.2 음악 기술 레이블 28

4.3 상대적 음고 기반 피아노 롤 인코딩에 관한 추가 실험 30
4.4 FLAGNet 30
4.5 결과 34
4.5.1 상대적 음고 기반 피아노 롤 인코딩에 관한 Ablation study . 34
4.5.2 FLAGNet 작곡 모델 평가 36

제 5 장 결 론 **41**
5.1 연구 의의 41
5.2 한계점 42
5.3 향후 연구 42

ABSTRACT **49**

감사의 글 **50**

표 차례

표 4.1 음악 기술 레이블 알고리즘 29
표 4.2 Validation Data에 대한 데이터 분류 성능 34
표 4.3 Mturk 설문조사 결과의 평균을 정리한 표. 40

그림 차례

그림 1.1 심볼릭 음악 중 하나인 악보. 음악의 정보, 각 악기의 정보, 그 악기가 연주에야할 음들의 정보가 수치화/도식화 되어 표현된 형태. 7

그림 1.2 심볼릭 음악을 Image로 처리하여 CNN을 활용해 새로운 음악을 생성해내는 MIDINet 모델. 9

그림 2.1 REMI 인코딩의 예시. 심볼릭 음악 데이터가 토큰화되어 저장된다. 13

그림 2.2 DAW에서 사용되는 피아노 롤의 예시. 본 그림은 음고가 일부만 표기되어 있다. 14

그림 2.3 CONLON 피아노 롤의 예시. 빠르게 반복되는 음과 음의 세기를 모두 반영할 수 있다. 18

그림 3.1 상대적 음고 피아노 롤 표현의 예시. 본 예시에서 n은 12이고 m은 8이다. 20

그림 3.2 Multi-vector conditional deep convolutional generative adverarial network의 생성기 구조 23

그림 3.3 FLAGNet 모델의 전체 구조 25

그림 4.1 MCDCGAN의 생성기 구현 31

그림 4.2 MCDCGAN의 판별기 구현 32

그림 4.3 클러스터링 결과 t-sne 시각화. 35

그림 4.4 VQVAE모델 Reconstruction 결과 36

그림 4.5 Mturk 설문조사 링크의 화면. 38

그림 4.6 Mturk 설문조사 결과의 Violin Plot. 39

제 1 장 서 론

1.1 연구 배경

심볼릭 음악은 귀로 들을 수 있는 형태의 WAV 음원 파일과는 다르게, 각 음의 시작점과 길이, 악기의 정보 등을 담고 있는 데이터를 의미한다. 현대 시대에 가장 널리 사용되는 형태는 MIDI(Musical Instrumnet Digital Interface)이며, 아래 그림1.1과 같은 악보 역시 가장 잘 알려져 있는 심볼릭 음악의 한 종류이다. 이런 심볼릭 음악은 음악 프로덕션, 음악 연주, 음악 편곡 등에 매우 유용하게 사용된다. 또한 일반적으로 상당히 용량이 높은 음원 파일과는 다르게, 비교적 가벼운 용량을 지니므로 음악 데이터를 저장하는 데에 있어 매우 효율적으로 사용된다.

MIDI [1]는 PC와 전자 악기간의 명령어와, 전자 악기로 보내는 연주용 음표의 형태, 그리고 악보를 구성하는 기록의 3가지 내용으로 구성된다. 이 중 사용자가 직접 접근하여 다루는 데이터는 일반적으로 악보를 구성하는 기록이다. 여기에는 하나의 음악에 해당 되는 Track이라는 단위가 존재하며, 그 안에 각 악기에 해당되는 Channel, 그리고 그 내부에 또 다시 각 음표의 음고, 길이, 세기 등의 정보가 담겨있는 형태로 이루어져 있다. 이러한 MIDI 데이터를 활용한 연구는 현대음악학, 실용음악, 정보과학 등 다양한 분야에서 이루어져왔다.

자동 음악 생성과 관련된 컴퓨팅 기술은 오랫동안 연구되어 왔다. 광범위한 음악 데이터 셋에 대한 접근성이 증가함에 따라 WAV 기반 기술, 심볼릭 음악 기반 기술 등을 기반으로 한 다양한 최신 모델들이 제안되었으며, 순환 신경망 [2], 적대적 생성 네트워크 [3], 강화 학습 [4], 트랜스포머 모델 [5] 등 다양한 딥 러닝 기술로 음악을 생산하는 모델이 연구되었다. 그러나 이들이 제안하는 기술은 미적인 음악을 제작

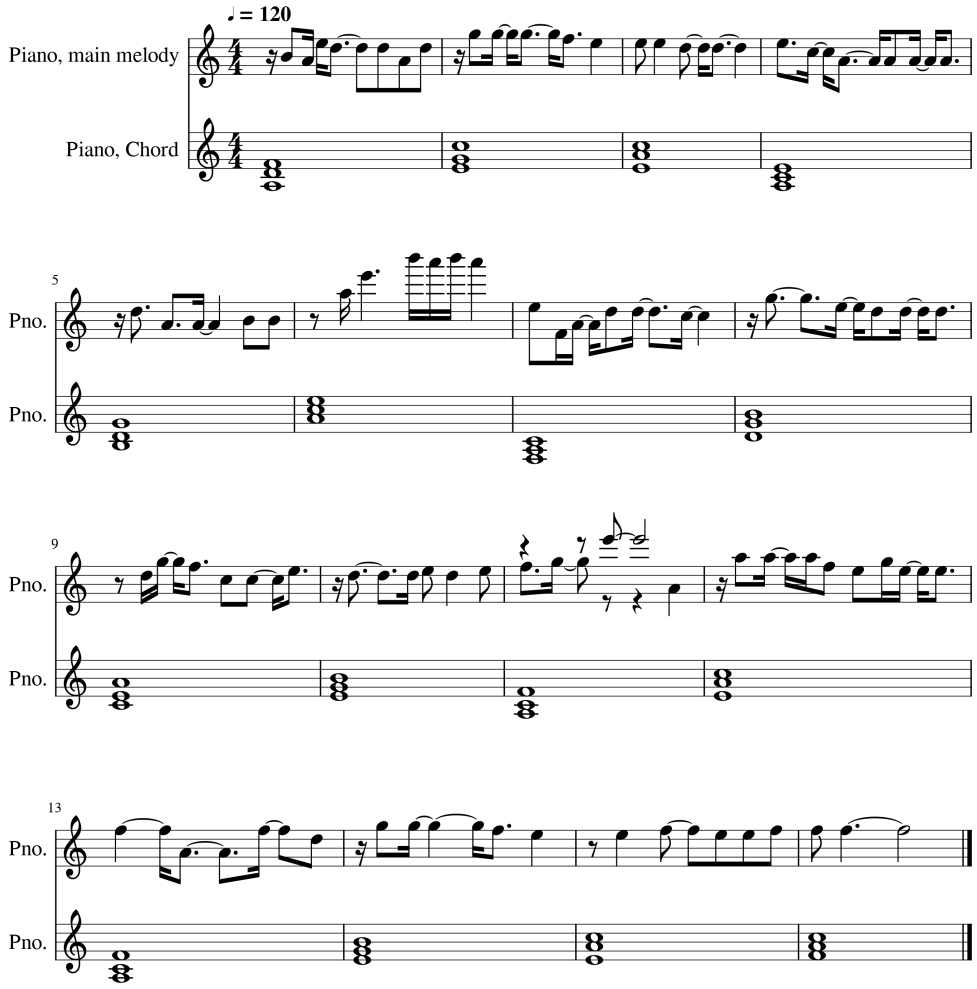


그림 1.1: 심볼릭 음악 중 하나인 악보. 음악의 정보, 각 악기의 정보, 그 악기가 연주에야할 음들의 정보가 수치화/도식화 되어 표현된 형태.

하는 것과는 다소 다르다. 기존의 자동 작곡은 종종 음표의 적합성과 데이터 셋과의 유사성 정도를 식별하여 모델을 구성하는 반면, 실제 인간의 작곡은 각 음표의 숙련된 응용을 통해 음악적 구조를 구성하고 이를 사용하여 음악 전체의 흐름을 만든다. 그러나 기존 생성 모델들은 음악에 대한 기반 지식을 적절히 활용하여 음악을 만들

어 내는 것을 고려하지 않거나, 이를 어려운 일로 받아들이는 경우가 많다.

음악적 기반 지식을 생성 모델에 적용하는 것을 목표로 하는 연구 역시 이루어지고 있다. 그 중에는 리듬, 음의 흐름(Contour) 및 음의 분리&통합과 같은 기반 지식을 분석하고 이를 심볼릭 음악 데이터인 MIDI(Musical Instrument Digital Interface) 데이터에 적용하여 생성하는 연구 [6]도 있으며, 또 다른 한 연구 [7] 에서는 Description-to-Sequence 작업으로 음악의 예술적 특징을 직접 글로 표현하면, 이를 하나의 Sequence로 encoding하여 제어하기도 하였다. 그리고 저수준의 음악적 특징을 모델링하여 고수준의 음악적 특징을 다룰 수 있도록 제안된 특성 모델링 연구도 이루어졌다. [8] 이러한 음악의 기반 지식을 인간 수준에서 적용하는 연구는 생성 모델의 확장성을 높일 뿐만 아니라 듣기 좋은 음악을 더 많이 만들어내는데 도움이 된다.

RNN과 트랜스포머와 같은 시퀀스 기반 모델들은 음악 생성을 위한 신경망 모델에서 가장 자주 사용되는 기술이다 [9] [5]. 이러한 경향이 생기는 큰 이유는 음악 역시 시간에 따라 음들이 배치되어 있는 시계열 데이터이기 때문이다. 그러나 음악은 단순한 시계열이라고 하기에는 입체적인 구조를 지니고 있어 MIDINet [10] 및 MuseGAN [3]모델과 같이 CNN, 특히 GAN을 사용하는 연구 또한 많이 이루어지고 있다. 그림2.1은 MIDINet의 심볼릭 음악 생성 구조를 나타낸다. 또한 음악의 스타일을 조건으로 주어 음악을 생성하기 위해 이미지 기반 모델인 CNN과 시계열 기반 모델인 Long Short Term Memory(LSTM)를 모두 사용하는 연구도 있다 [11].그들은 음악 생성 알고리즘에 CNN을 사용하는 것이 꽤 유효하고 학습 시간에 대해 더 나은 성능을 가질 수도 있음을 입증했다.

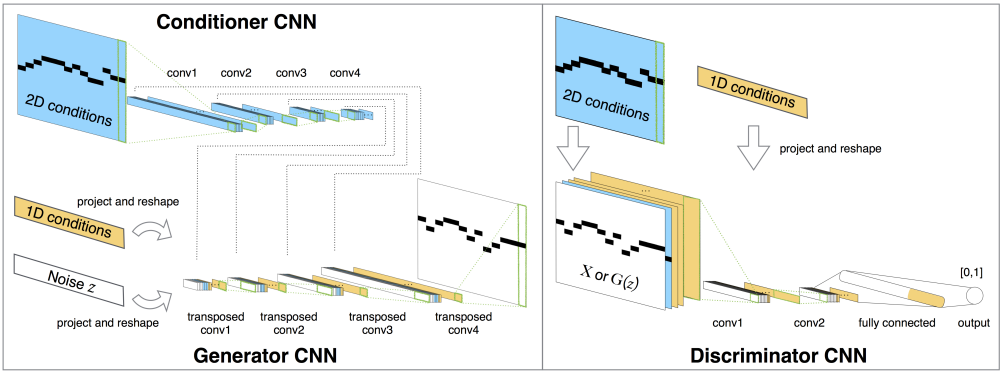


그림 1.2: 심볼릭 음악을 Image로 처리하여 CNN을 활용해 새로운 음악을 생성해내는 MIDINet 모델.

1.2 연구 목표

이러한 배경을 바탕으로, 음악적 기반 지식을 생성 모델에 적용하기 위해 생각해낸 아이디어는 마디 수준에서 음악적 기술을 레이블링하고 Deep Conditional GAN(DCGAN) [12]을 사용하여 그 레이블들을 Condition으로 주어 MIDI가 생성될 수 있는 환경을 구성하는 것이다. 그러나 마디 수준의 레이블을 DCGAN으로 구조에 삽입하여 MIDI를 생성하는 것만으로 음악을 만드는 데는 한계가 있다. 왜냐하면 음악을 구성함에 있어 따라 다른 마디들의 흐름을 따라가는 과정이 필요하기 때문이다. 따라서 우리는 2가지 방법을 통해 생성되는 마디가 음들의 흐름을 따를 수 있도록 한다. 하나는 음악 데이터 셋에 주어진 각 마디의 음악적 기술 레이블들을 기반으로 Gated Recurrent Unit(GRU)와 LSTM 구조가 포함되어 있는 RNN 모델을 학습시켜 음악 기술 레이블의 시퀀스가 어떻게 조합되어야 하는지를 학습한다. 이를 통해 어떤 음악적 기술 레이블을 생성해야 자연스러운 마디를 생성할 수 있는지를 표현해줄 수 있다. 또한 주어진 레이블을 기반하여 생성하는 과정에서 레이블 벡터만을 주는 것이 아니라 선제 되는 마디의 음들을 벡터화하여 Condition을 주게 된다. 이 때 pix2pix [13]모델의 아이디어를 사용하여 선제 되는 마디의 음들에 기반

해 MIDI 생성이 조절될 수 있도록 한다.

이미지 기반 분류 모델과 생성 모델을 사용하기 위해, 우리는 피아노 롤 기반 인코딩 방식으로 MIDI 데이터를 처리한다. 피아노 롤 기반 인코딩 방식은 심볼릭 음악 연구 분야에서 이미지 기반 인코딩 방식의 절대적인 비중을 차지하고 있는 방식으로, 주어진 심볼릭 음악을 하나의 이미지 상에 표현하는 방식이다. 피아노 롤 기반 인코딩 방식은 특히 심볼릭 음악 데이터 표현을 위해 많은 연구 [14] [**?**]에서 사용되었다. .

이러한 피아노 롤 기반 인코딩 방법은 이미지 기반 모델을 쉽게 사용할 수 있고 시각적으로 직관적인 정보를 제공할 수 있다는 장점이 있다. 그러나 기존의 피아노 롤 기반 인코딩 방법은 사용 가능한 모든 음고(Pitch)를 제어하면서 데이터의 크기가 많이 증가하고, 데이터가 희소(Sparse)해진다는 단점이 있다.

따라서 우리는 이미지 크기를 적절하게 줄일 수 있도록 상대적 음고 변화를 기반으로 한 피아노 롤 기반 인코딩 방법을 제안한다. 이 접근 방식은 데이터 셋에서 주어진 음악의 음계에 의존하지 않고 학습과정에서 음악 데이터를 사용할 수 있도록 하며, 생성된 마디를 12개의 기본 음계(C, C#, D, D#, E, F, F#, G, G#, A, A#, B) 및 각 마디의 코드와 매칭하여 사용할 수 있기 때문에 생성된 음악에 대한 더 유연한 활용 범위를 제공한다. 또한 MIDI 마디를 상대적인 음높이 변화로 처리하여 절대적인 음고 값은 잃게 되지만 마디에 주어진 음악적 기술을 유지할 수 있는데, 이는 위 인코딩 방식에서 MIDI 마디가 음의 흐름 정보를 유지하기 때문이다.

결과 모델인 FLAGNet은 음악 바에 포함된 음악 기술을 사용하여 음악 영역 지식을 이해하고, 음악 기술의 순서를 분석하여 음악의 전체적인 흐름을 제어하고, 생성된 이미지를 처리하여 심볼릭 음악을 만들 수 있을 뿐만 아니라 모든 기본 음계

및 마디의 코드를 활용할 수 있다.

제 2 장 배경 이론 및 관련 연구

본 장에서는 연구에 필요한 배경 이론들과 진행된 연구에 대한 관련 연구를 기술한다.

2.1 배경 이론

2.1.1 심볼릭 음악의 표현 방식

심볼릭 음악을 사용하는 연구에 있어, 가장 중요하게 다루어지는 것 중 하나는 심볼릭 음악을 어떤 형태로 인코딩하여 사용할 것이냐는 점이다. 심볼릭 음악 그 자체를 사용하기 보단, 심볼릭 음악에 담겨있는 정보를 정제하여 전처리를 한 이후에 사용하는 것이 일반적으로 많이 사용되는 형태이다.

이러한 심볼릭 음악을 인코딩하는 방법중에 가장 많이 사용되는 두 가지가 사건 기반(event-based) 방식 [15] [16] [5]과 이미지 기반(image-based) [3] 방식이다.

첫째로, 사건 기반 방식에서는 심볼릭 음악에 담겨있는 다양한 정보들이 시계열 토큰으로 변환되어 사용된다. 대표적인 방식으로는 REMI [17]가 있다. REMI는 각 음들의 정보 하나하나를 토큰화(Tokenization) 하여 데이터를 구성하는 형태이다. 이런 형태의 인코딩 방식은 심볼릭 음악을 길이가 긴 하나의 1차원 시계열로 나타내기 때문에, 트랜스포머 [18]와 같은 고성능의 시계열 분석 모델을 활용하기 위해 사용되는 경우가 많다.

두 번째로, 이미지 기반 방식에서는 심볼릭 음악에 담겨있는 정보들이 하나의 이미지로 표현되어 사용된다. 가장 널리 사용되는 방식은 피아노 롤이 있다. 고전적인 피아노 롤은 피아노의 연주를 위하여 피아노의 악보를 하나의 좌표체계 하에 표현

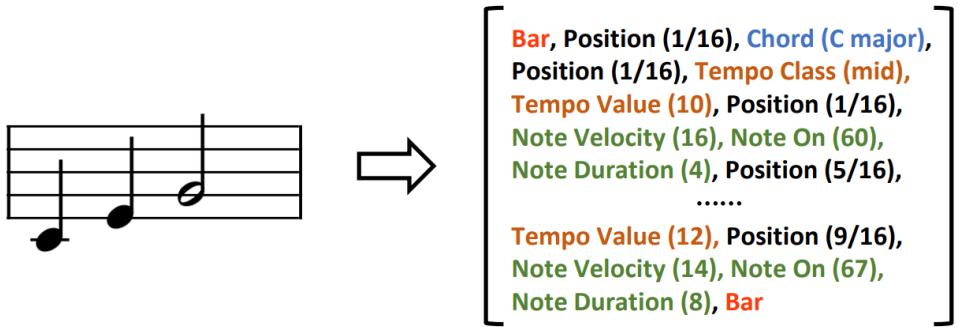


그림 2.1: REMI 인코딩의 예시. 심볼릭 음악 데이터가 토큰화되어 저장된다.

한 형태로 이루어져 있었으며, 이러한 형태가 아이디어가 현대사회로 이어져와서 DAW(Digital Audio Workstation)에서 악보, 즉 심볼릭 음악을 시각화하기 위해 사용되는 일반적인 방식으로 자리잡았다. 그림2.2에 널리 사용되는 형태의 피아노 롤 예시가 있다. 일반적인 피아노 롤은 가로축은 시간, 세로축은 음의 높이를 의미하며, 특정 음이 연주되는 시간에, 그 음의 높이에 해당되는 부분에 1을, 그리고 나머지 부분에 0을 주는 이진 행렬의 형태로 사용한다. 이러한 형태의 인코딩 방식은 입체적인 구조의 파악과 VQVAE와 같은 다양한 이미지 기반 모델을 활용하기 위해 사용되는 경우들이 많다. [19] [20]

2.1.2 조건부 적대적 생성 네트워크

적대적 생성 네트워크(Generative Adversarial Networks, GAN) [21]은 생성자 G와 판별자 D라는 두 개의 적대적인 모델을 포함하고 있다. 생성자 G는 주어진 데이터들의 분포를 학습하여, 저차원의 노이즈 $z \in \mathbb{R}$로 부터 주어진 데이터와 유사한 새로운 데이터를 생성해내는 것을 학습한다. 판별자 D는 주어진 데이터에 대해서, 이것이 정말 주어진 데이터인지, G가 생성해낸 가짜 데이터인지 구별해내는 것을

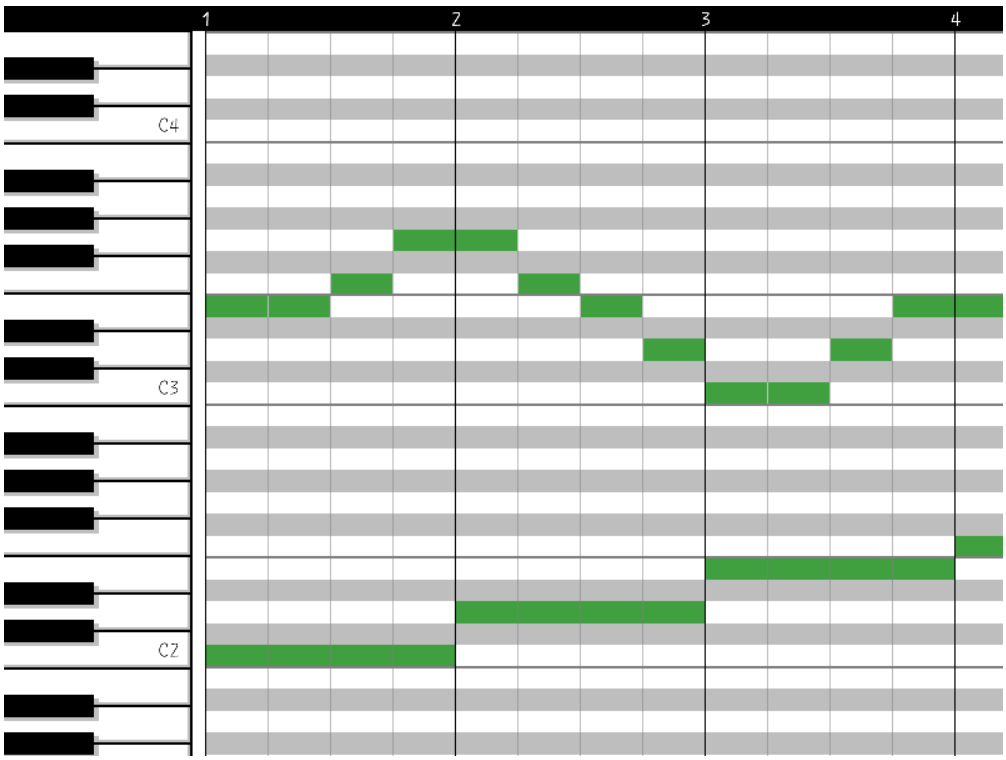


그림 2.2: DAW에서 사용되는 피아노 롤의 예시. 본 그림은 음고가 일부만 표기되어 있다.

학습한다. 주어진 데이터 X에 대해서, G와 D는 아래 식에 기반하여 학습을 진행하게 된다.

$$\begin{aligned}\min_G \max_D V(G,D) = \mathbb{E}_{x\sim P_X}[\log(D(x))] \\ + \mathbb{E}_{z\sim P_Z}[\log(1-D(G(z)))]\end{aligned} \tag{2.1}$$

$x \sim P_X$ 는 실제 데이터의 분포를 의미하고, $z \sim P_Z$는 무작위 노이즈 데이터의 분포를 의미한다.

하지만 단순히 데이터를 생성해내는 것 뿐만아니라, GAN 모델에 특정 레이블을 조건으로 주어 그에 맞게 생성을 학습하도록 할 수도 있다. 조건부(conditional) GAN [**?**]에서는, 조건 $c \in \{f_1, f_2, \ldots, f_n\}$이 GAN 모델에 추가된다. G 와 D는 주어진 조건 y에 대하여 학습이 된다. 이러한 상황에서의 모델이 학습하게 되는 식은 아래와 같으며, y가 단순히 GAN의 목적 함수에 추가된 형태이다.

$$\begin{aligned}\min_G \max_D V_c(G,D) = \mathbb{E}_{x\sim P_X}[\log(D(x|c))] \\ + \mathbb{E}_{z\sim P_Z}[\log(1-D(G(z|c)))]\end{aligned} \tag{2.2}$$

이는 G가 주어진 특정 레이블, 즉 조건 y에 대하여 데이터를 생성해내는 것을 학습할 수 있도록 한다. 위와 같이 레이블을 주어 GAN Model에 조건을 추가하는 방식은 조건이 결국 특정 레이블로 주어져 있어야 한다는 점과 조건을 하나의 Singleton 값으로밖에 할 수 없다는 점이 한계가 있다.

Pix2Pix [22] 모델에서는 이러한 조건을 하나의 값이 아닌 이미지, 즉 고차원 데이터로 조건 c를 주어 목적하는 Target x_t을 만들 수 있도록 하는 모델을 제안한다. 이 경우 목적함수는 2가지 항으로 나누어지는데, GAN에서 사용되던 데이터의 분포를 따르도록 하는 1가지 항과, 주어진 Condition에 대하여 그 Target을 얼마나 잘 따라해 내었는지를 확인하는 L1 손실함수를 추가하여 더해주는 식으로 구성되어

있다. 목적 함수는 다음과 같다.

$$
\begin{aligned}
G^* &= \min_G \max_D [V(G,D) + \lambda L_{L1}(G)], \text{where} \\
L_{L1}(G) &= \mathbb{E}_{x,c,z} \| x_t - G(z|c) \|_1
\end{aligned}
\tag{2.3}
$$

위와 같은 방식을 활용하여 다차원의 데이터를 조건으로 주어, 이에 맞게 새로운 데이터를 생성해낼 수 있도록 학습할 수 있다.

2.2 관련 연구

2.2.1 자동 음악 작곡

자동으로 음악을 작곡하는 모델은 오랫동안 연구되어 왔지만, 특히나 다양한 딥러닝 모델의 발전 하에 그 성능이 비약적으로 상승해왔다. 그 중에는 Wav 데이터를 직접 학습하여 음악을 생성하는 JukeBox [23]와 같은 모델이 있고, 심볼릭 음악 데이터들을 활용하여 음악을 생성하는 모델들이 있다. 심볼릭 음악 데이터들을 활용하여 음악을 생성하는 모델은 또 다시 이미지 기반의 인코딩을 활용하는 연구와 사건 기반의 인코딩을 활용하는 연구로 나누어 진다.
사건 기반의 인코딩을 활용하는 연구에는 Music Transformer [5], POP Music Transformer [17] 등이 있으며, 이미지 기반의 인코딩을 활용하는 연구에는 MIDINet [10], Symbolic Loop 생성 모델 [19] 등이 있다. 특히나 최근에는 Transformer 모델을 활용하여 데이터를 분석하는 다양한 모델들이 최첨단의 성능을 지니고 있다.

이들 중에서도 음악적 기반 지식을 생성 모델에 적용하는 것을 목표로 하는 연구 역시 이루어지고 있다. 그 중에 한 연구 [6]는 리듬, 음의 흐름(Contour) 및 음의 분리&통합과 같은 기반 지식을 분석하기 위해 Extraction-Residual Latent Space Decoupling Model algorithm을 활용하여 직접 음악적 요소들을 추출해낸다. 그 후에 LSTM 기반 모델을 활용하여 심볼릭 음악 데이터인 MIDI(Musical Instrument Digital Interface) 데이터에 적용한다. 또한 MuseMorphose [24]라는 연구에서는 하

나의 Transformer에 In-attention conditioning 기술을 적용하여 음악적 기반 지식을 별도로 넣어주고, 이에 맞게 음악을 생성하는 결과를 보여주었다.

2.2.2 피아노 롤의 새로운 표현 방식

피아노 롤 기반 인코딩 방법은 이미지 기반 모델을 활용할 수 있다는 점에서 장점을 지니지만, 동시에 많은 단점들이 존재한다. 첫째로, 데이터가 굉장히 희소(Sparse)해진다. 모든 시간과 모든 음고를 전부 담을 수 있는 격자(Grid)의 형태의 이미지에 연주 되는 음 만의 정보가 들어가기 때문에 들어가는 정보의 양에 비하여 데이터의 크기가 커지게 되고, 이러한 현상을 데이터의 용량이 단순히 늘어나는 것 뿐만 아니라 모델을 학습시키는 과정에서 성능을 떨어트리는 이유가 될 수 있다. 둘째로, 음의 세기 정보를 담을 수 없고, 반복되는 음과 한번 길게 연주하는 음을 구분할 수 없다. 셋째로, 음의 높이를 이미지의 높이 값으로 구분하는 형태의 데이터이기 때문에 음악의 화음이나 코드, 차트 등의 정보를 반영하기 어렵다.

이들 중 두 번째 단점을 해결하기 위해 제안된 새로운 형태의 피아노 롤이 있다. 이는 CONLON 피아노 롤으로, 연주되는 음의 세기와 길이를 담는 2개의 Channel을 구성하여 피아노롤을 발전시킨 형태이다. 그림2.3에 CONLON 피아노 롤의 예시가 나와있다. 하지만 이러한 형태를 사용하고도 데이터는 여전히 희소하다는 단점이 존재했고, 음악의 화음과 코드, 차트 등의 정보를 다루기 어렵다는 단점은 그대로 존재하였다.

우리는 이런 단점을 해소하기 위하여 상대적 음고의 변화만을 담는 인코딩 방식을 제안한다. 또한 우리가 제안하는 피아노 롤은 일반적인 방식의 피아노 롤에 적용하기 때문에 음의 세기 정보를 담을 수 없고, 반복되는 음과 한번 연주되는 음을 구분할 수 없다는 단점이 존재할 수 있다. 하지만 우리가 제안하는 인코딩 방식의 경우 음의 세기 정보를 담아서 사용하기 때문에 음의 세기 정보를 담을 수 있으며, CONLON 피아노롤에서의 확장 역시 가능하므로 이와 같은 단점들을 제거할 수 있다.

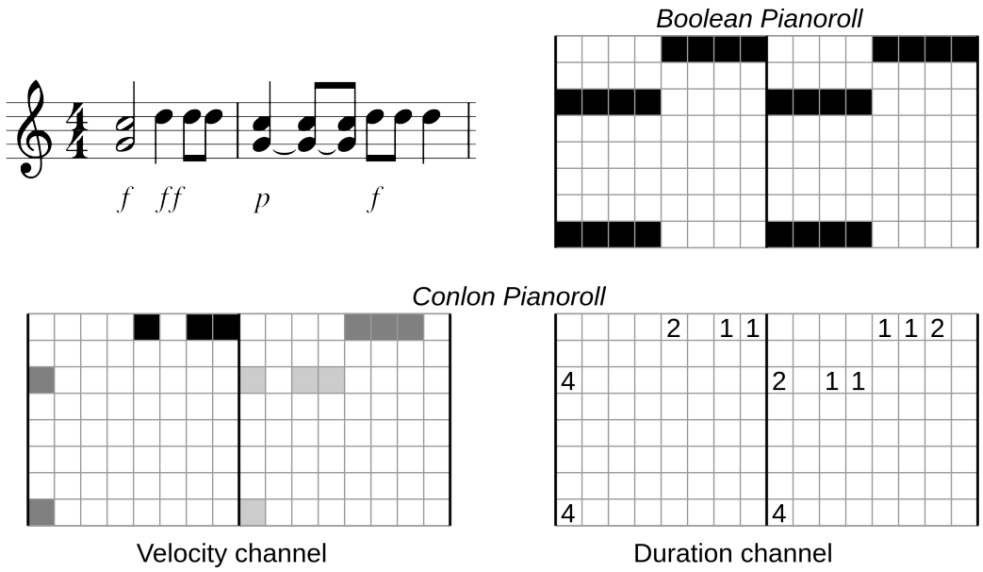


그림 2.3: CONLON 피아노 롤의 예시. 빠르게 반복되는 음과 음의 세기를 모두 반영할 수 있다.

제 3 장 제안 기법

이 장에서는 자동 음악 생성 모델인 FLAGNet을 구성하기 위하여 사용한 MIDI의 전후처리 방식과 구성한 생성 모델에 대하여 설명한다.

3.1 MIDI의 상대적 음고 피아노 롤 표현

MIDI 데이터를 행렬로 표현하는 방식인 피아노 롤 기반 인코딩 방식은 음악 생성 연구에 널리 사용되어 왔다. [14] 우리는 이러한 인코딩 방식을 기본으로 하되 피치 정보를 상대적인 값으로 처리하여 최적화하는 상대 음고 피아노 롤(Relative Pitch Pianoroll)을 제안한다. 우리는 단지 상대적인 차이로만 피치 정보를 유지한다. 이 방법은 음의 흐름 정보를 유지하면서 데이터의 크기를 줄이고 피아노 롤 매트릭스가 너무 희소(Sparse)해지는 것을 방지할 수 있다. 그리고 우리는 디코딩 과정에서 어떤 피치를 사용할지 결정하기만 하면 어렵지 않게 MIDI 데이터를 얻어낼 수 있다. MIDI의 각 마디는 행렬 $M \in \{0, V \in [1, 128]\}^{n \times m}$로 표현된다. 여기서 V는 음의 세기(Velocity)를 의미하고 n은 사용 가능한 피치 변경 범위이며 m은 하나의 마디에 존재하는 타임스탬프를 나타낸다. 행렬 M의 열 변화는 시간의 변화를 의미하며, 행 변화는 피치의 변화를 의미한다. 각 매트릭스에서 가장 낮은 피치를 가진 음을 가장 낮은 행에 일치시킴으로써, 우리는 더 적은 수의 피치를 사용하여 마디의 음악적 특성과 마디의 음고 변경 정보를 잃지 않으면서 행렬 크기를 작게 최적화할 수 있다. 우리는 MIDI 처리 중에 주어진 음계나 마디의 코드에 행렬을 매칭하여 MIDI를 생성한다. 우리는 이 방식을 통해 피아노 롤과 같이 MIDI를 이미지로 사용하고 연구에 이미지 기반 기술을 적용할 수 있다. 또한 우리는 처리한 데이터를 보는 것만으로 음악의 흐름을 직관적으로 파악할 수 있다. 에 위 방식의 간단한 예시가 제안되어

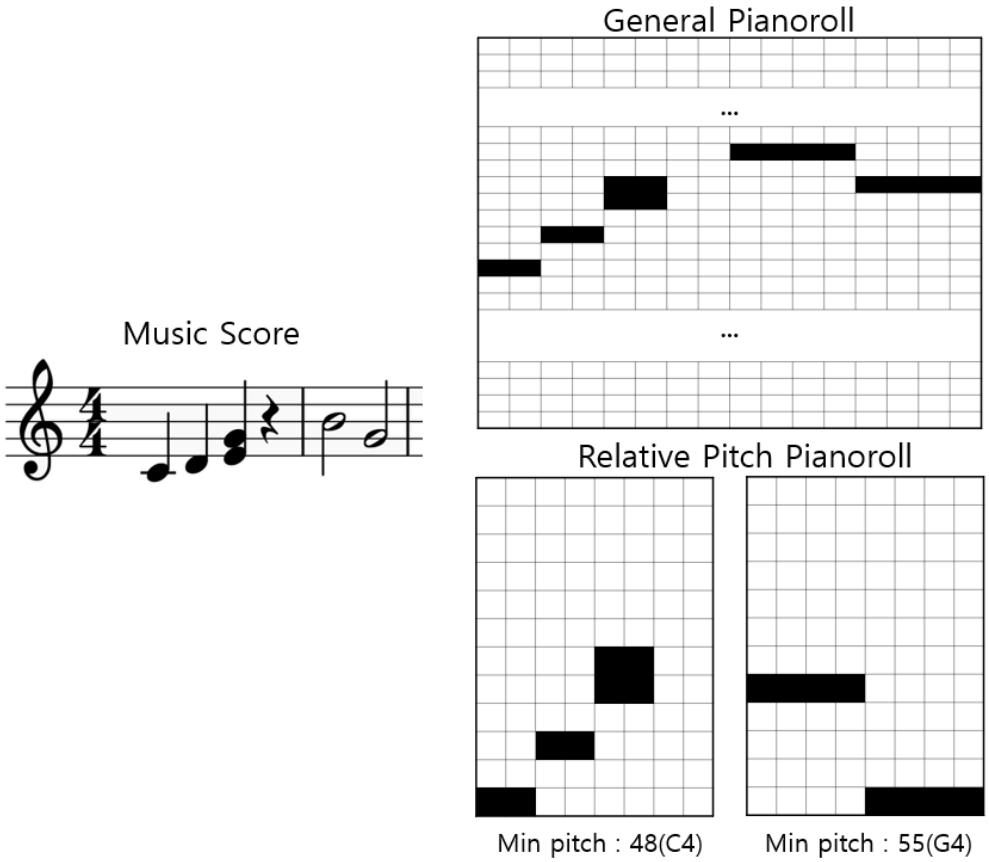


그림 3.1: 상대적 음고 피아노 롤 표현의 예시. 본 예시에서 n은 12이고 m은 8이다.

있다.

3.2 다중 벡터 조건부 딥 콘볼루션 적대적 생성 망

적대적 생성 네트워크 [25]에는 생성 모델 G와 판별 모델 D라는 두 가지 적대적 모델이 포함되어 있다. 생성 모델 G는 주어진 데이터의 분포를 학습하고, 이를 기반으로 저차원 노이즈 z∈ R 에서 주어진 데이터와 유사한 구조로 데이터를 생성하는 방법을 학습한다. 판별기 모델 D는 주어진 데이터가 생성 모델 G에서 생성되는지 또는 실제 데이터로부터 생성되는지 구별하는 형태로 학습을 진행한다.

조건부 적대적 생성 네트워크에서 조건 $c \in \{f_1, f_2, \ldots, f_n\}$가 GAN 모델에 추가된다. G와 D는 주어진 레이블 c에 대해 훈련될 것이다. 이를 통해 생성기 G는 주어진 레이블 $f_{i[1,n]}$에 해당하는 MIDI 이미지를 생성할 수 있다. 심층 컨볼루션 GAN 구조에서, 생성기 G와 판별기 D의 구조는 컨볼루션 신경망으로 구성된다. 우리는 또한 레이블을 입력 노이즈에 연결하고 위의 조건부 GAN 구조를 가진 상태에서 조건부 DCGAN을 구현할 수 있다.

그러나 음악의 마디는 단순히 음악적 기술 레이블 정보만 가지고 있는 것이 아니라 주변 마디의 내용과 조화를 이루어야 한다. Pix2Pix [22]의 아이디어에 근거하여, 우리는 학습 이미지 생성기에 대한 또 다른 손실 함수를 추가하여 다중 벡터 조건부 딥 컨볼루셔널 적대적 생성 망을 구성하였다.(Multi-vector Conditional Deep Convoluational Generative Adversarial GAN, MCDCGAN) 추가되는 벡터 조건은 선제 되는 음들의 조건 c_{pr} 및 해당 타겟 이미지 x_t이다. 실제 학습에서는 Relational 피아노 롤 형태의 대상 이미지를 설정하고 선제 되는 음들의 상대 시간 변화 및 피치 변화를 사용하여 조건 벡터c_{pr}를 설정한다. 대상 이미지의 바로 직전에 등장하는 음의 음고를 기준으로 음들의 상대적인 음고 값들을 하나의 벡터로 구성하고, 타겟 이미지의 시작점을 0초로 하여 상대적인 시간 값들을 하나의 벡터로 구성한다. c_{pr}는 두 개의 선형 벡터, 즉 상대 시간 변경 정보 벡터와 상대 피치 변경 정보 벡터로 구

성된다. 조건 벡터가 주어지면 타겟 이미지와 같은 이미지를 생성하는 것을 목표로 하는 L1 손실 함수를 추가한다. 그래서 MCDCGAN의 최종 목표 함수는

$$
\begin{aligned}
L^* &= \min_G \max_D [V_c(G, D) + \lambda L_{L1}(G)], \text{where} \\
L_{L1}(G) &= \mathbb{E}_{x,y,z} \| x_t - G(z|c_{pr}, c) \|_1
\end{aligned}
\tag{3.1}
$$

의 형태가 된다. V_c는 조건부 GAN의 손실함수를 의미하며, 값은 선제 되는 음들에 관한 Condition인 c_{pr}으로 인한 L1 loss값의 균형을 맞추기 위해 사용되는 Parameter이다.

3.2.1 모델 입력 및 출력

생성기 모델 G의 경우 노이즈, 마디에 해당하는 음악적 기술 레이블 조건, 선제되는 음표의 시간 조건, 선제되는 음들의 음고 조건이 포함된다. 노이즈는 길이 100의가우시안 노이즈를 사용하고, 마디에 해당하는 음악적 기술 레이블 조건은 원-핫 인코딩 된 단일 레이블 조건이 들어가게 된다. 선제 되는 음표의 시간 조건과 선제되는 음들의 음고 조건은 앞에서 사용된 음들의 상대적인 타이밍 값과 상대적인 음고 값들을 길이 32의 1차원 시계열로 배치되어 사용된다. 데이터의 수가 32개보다 적은 경우 패딩되어 사용된다. 위에 기반한 생성기 모델 G 는 그림 3.2와 같이 표현된다.

판멸기 모델 D의 경우 생성기 모델 G와 유사하지만, 노이즈 대신에 실제 이미지 값이 주어지게 된다.

3.3 MIDI 후처리

생성된 Image를 MIDI로 처리하는 과정이 필요하다. 우선 생성된 행렬 M을 최소 요소가 0이고 최대 요소가 1이 되도록 정규화 한다. 그 후 임계 값보다 큰 값을 가진

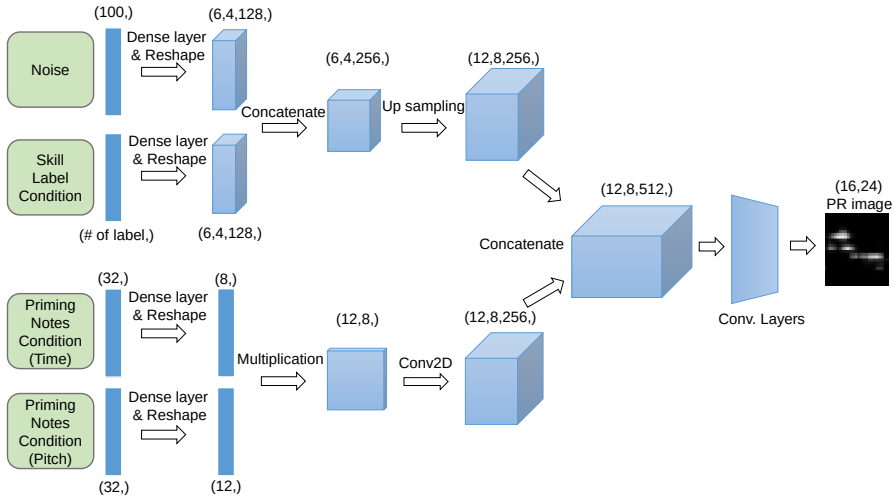


그림 3.2: Multi-vector conditional deep convolutional generative adverarial network의 생성기 구조

일부 요소만 남기고 나머지는 0으로 변경한다. 그런 다음 Local Maximum 필터를 사용하여 피크 행렬 $M_{P_{I,j}}$을 구한다. M의 3×3부분행렬인 행렬 $M_{[i-1,i+1][j-1,j+1]}$를 C_{ij}라고 하자. $i-1, i+1, j-1, j+1$이 행렬 크기를 벗어나는 경우 2행 또는 2열만 고려한다. Local Maximum 필터로 피크 행렬은 다음과 같이 나타낼 수 있다.

$$M_{P_{i,j}} = \begin{cases} 1, & \text{if } \max(C_{ij}) = M_{i,j} \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

각 피크, 즉 $M_{P_{I,j}}$에서 값이 1이 되는 위치를 각 음들의 위치로 사용한다. 각 피크위치에 대하여 행렬M 에서의 값들을 확인하여 음들의 시작 위치와 음들의 연주시간(Duration)을 결정한다. 특정 임계 값을 정해두고 피크 위치에서 오른쪽 값들을 확인하여 임계 값보다 값이 내려가는 순간 까지를 음의 길이로 지정할 수 있다. 음의 시작점 역시 피크 위치를 그대로 사용하는 것 보다 피크 위치에서 왼쪽 값들을 확인하여 사용한다. Pianoroll에서의 음의 시작점은 각 음을 표현하는 부분의 중간

지점이 아닌 가장 좌측 지점이기 때문이다. 그 후에 각 피크 위치의 행렬 M에서의 값을 음의 세기(Velocity)로 결정한다. 이러한 방식으로 생성자 G에 의해 생성된 행렬 이미지는 이진 행렬인 피아노 롤의 형태로 다시 표현될 수 있으며, 심볼릭 음악 데이터인 MIDI로 나타낼 수 있다.

하지만 심볼릭 음악 데이터인 MIDI로 나타내기 위해서는 음고의 절대적 위치를 결정해줄 필요가 있다. 이를 위해 가능 음고 집합을 정의하게 되는데, 가능 음고 집합은 기준 음고를 기준으로 12개의 음고를 순차적으로 선택하여 하나의 집합으로 지정하게 된다. 집합의 크기가 12인 이유는 하나의 옥타브가 12개의 반음으로 구성되어 있으므로 12개의 음을 다루어야 모든 음계와 코드를 가장 잘 맞추는 멜로디를 구할 수 있기 때문이다.

여기서 기준 음고는 이전 마디의 마지막 음의 높이이고, 만약 첫 마디라면 기준 음고는 48(C4)로 지정한다. 기준 음고를 기준으로 12개의 음고를 결정하는 과정에서 가능 음고 집합을 만드는 방향을 올려서 결정할 수도 있고 내려서 결정할 수도 있는데, 이는 현재 주어진 마디의 Relational Pitch 피아노 롤을 이진 분류기에 학습시켜 얻어낸다.

첫 번째 음이 결정되면, 우리는 마디에 있는 나머지 음의 음고를 쉽게 결정할 수 있다. 예를 들어, 피크 행렬에서 나타나는 한 마디에 있는 음들의 상대 피치값이 [1,6,3]이고 기준 음고가 48, 가능 음고 집합을 만드는 방향이 음이 올라가는 방향이라면, 사용 가능한 음의 음고 집합은 [48, 53, 50], [49, 54, 51], . . . [59, 64, 61]이 될 것이다. 주어진 마디의 코드(Chord)와 음악의 음계를 기반으로, 가장 음계와 코드에 잘 맞는 집함을 최종적인 절대 음고로 사용한다. 음계와 코드가 주어져 있다면 그에 맞는 음고 집합을 구해낼 수 있는데, 이를 사용 가능한 음고 집합들과 직접 비교하여 가장 많은 음이 조화 음고 집합에 속하는 경우를 선택한다.

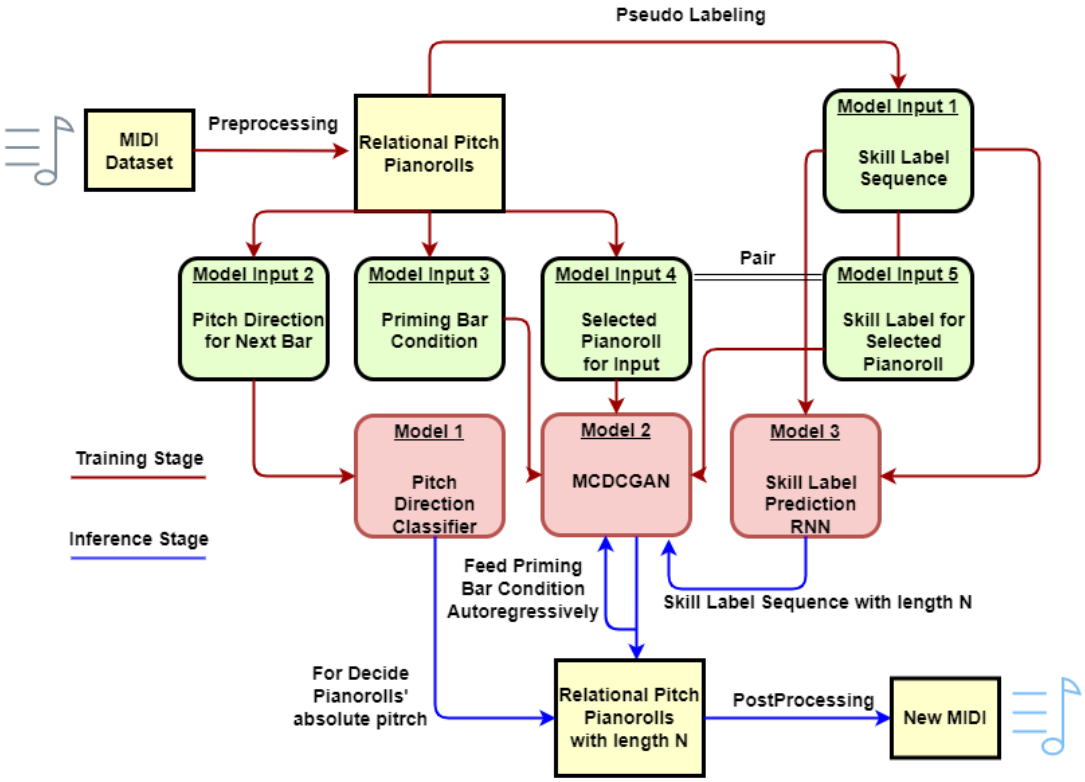


그림 3.3: FLAGNet 모델의 전체 구조

3.4 FLAGNet의 전체 구성

FLAGNet의 기본 구조는 그림 3.3에 나와있다.

우리는 먼저 주어진 MIDI 데이터 셋을 상대 음고 피아노 롤 형태로 처리하고 MIDI의 음의 흐름 정보에 따라 해당 기술 레이블과 음고 방향 레이블을 부착한다.

기술 레이블은 주어진 마디 행렬에 대한 인간의 음악적 스킬 정보를 포함한다. 우리의 지식 하에, 각 마디에 음악 기술 레이블에 대한 정보가 있는 데이터 세트는 없다. 그래서 우리는 휴리스틱 음악 알고리즘이나 비지도 방식(Unsupervised Method)으로 레이블을 정의해야 한다. 그러나 비지도 방법에서 클러스터링과 같은 계산 알고리즘으로 기술 레이블을 얻는 것은 인간의 음악 기반 지식을 반영한다는 의도에서 벗어나게 된다. 그래서 우리는 음의 흐름 정보를 분석하여 휴리스틱 알고리즘으로 Pseudo-레이블을 정의하고 이 Pseudo-레이블을 사용하여 다중 레이블 분류기를 훈련시켜 다시 주어진 데이터를 분류할 수 있도록 하였다. 이러한 형태는 본 휴리스틱 알고리즘처럼 레이블이 다소 Noisy할 수 있다고 판단되는 상황에서 Self-Ensembling효과를 주어 Wrong Label을 Filtering할 수 있다. [26] 휴리스틱 알고리듬이 어떻게 구성되어 있는지는 제 3장에서 후술한다. 이 분류기를 사용하면 휴리스틱 방법에서 생기게 되는 일부 오류를 피할 수 있으며, 하나의 마디마다 하나의 음악적 기술 레이블을 가지고 있는 형태로 데이터를 준비할 수 있다. 방향 레이블은 바로 다음 마디에서 절대 음고를 결정할 때에 사용된다. 음악 데이터 셋에는 마디들의 시퀀스가 포함되어 있으므로 각 막대에 대한 스킬 레이블을 정의할 수 있고, 이에 따른 각 음악에 대한 기술 레이블 시퀀스를 찾을 수 있으며, 이러한 레이블 시퀀스는 순환 신경망(RNN) 모델의 학습에 사용된다. 정리하면, RNN은 기술 레이블을 예측하고 그 예측한 레이블에 맞게 Multi-vector Conditioned Deep Convolutional GAN은 기술 레이블과 선제 되는 마디들의 정보로 피아노 롤 이미지를 생성한다. 최종적으로, 사용자의 음악적 요소 입력을 고려하여 MIDI를 만드는 처리를 하여 완성된 음악을 생성해낼 수 있다.

제 4 장 실험

이 장에서는 데이터셋과 이를 전후처리 하는 과정, 그리고 이를 모델에서 어떻게 활용하는지를 설명한다.

4.1 데이터셋

우리는 POP909 데이터 셋을 FLAGNet 학습에 사용했다. POP909는 전문 뮤지션들이 창작한 909곡의 대중가요 피아노 편곡 버전을 수록한 데이터 셋이다. 데이터 세트의 구성은 보컬 멜로디, 리드 악기 멜로디, 그리고 각 곡에 대한 피아노 반주를 MIDI 형식으로 포함하고 있으며, 이들은 원본 오디오 파일에 맞게 정리되어 있다. [27]

데이터 다양성을 위해 FLAGNet 모델 학습의 경우 MIREX 2019 [28]의 Patterns for Prediction Development Dataset(PPDD)을 추가로 사용했다. 하나의 모델에 두 개의 데이터를 전부 사용한 것이 아니고, 각 데이터에 대하여 모델을 각각 학습시켜 사용하였다. 이 데이터는 Lakh 데이터 집합 [29]의 하위 집합으로, 비교적 단순하고 다루기 쉬운 구조를 가진 심볼릭 음악을 많이 포함하고 있기 때문에 심볼릭 음악 관련 연구에서 널리 사용된다. PPDD 데이터 세트의 가장 큰 버전은 10000개의 단선율 음악 MIDI와 10000개의 다성 음악 MIDI를 포함한다.

우리는 이 데이터 셋에서 멜로디만 사용했고 4/4 박자를 지닌 데이터만 사용했다. 다른 박자들을 사용하는 것도 가능하지만, 피아노 롤 행렬기반 모델에서 다른 박자를 지닌 데이터들을 똑같은 사이즈의 피아노 롤로 표현하는 과정에서 데이터의 구조를 충돌시켜 모델의 성능을 저하시킬 수 있기 때문에 4/4 박자를 지닌 음악만

사용했다.

4.2 음악 기술 레이블

MIDI 데이터에는 마디의 음악적 기술을 직접 설명할 수 있는 라벨이 없기 때문에 마디별로 스킬 라벨을 따로 정할 필요가 있다. 따라서 휴리스틱 알고리듬을 만들어 MIDI 데이터를 사용하여 얻은 통계 정보를 활용하여 각 마디에 스킬 레이블을 붙였다. 먼저 MIDI 데이터를 기반으로 각 마디에 대한 음고 변경 정보, 음 지속 시간 정보, 음 타이밍 정보 및 휴식 정보를 수집했다. 그런 다음 이 정보를 사용하여 휴리스틱 알고리즘을 기반으로 각 막대에 레이블을 부착했다. 자세한 알고리즘은 표4.1와 같다. n값은 지정할 수 있는 변수로, 실제 실험에서는 모든 레이블에 대해 66

휴리스틱 알고리듬에 의한 레이블을 직접 사용하여 일부 잘못된 특이 케이스를 처리하는 것은 어렵다. 위 알고리즘에 의해 처리되었다고 하더라도, 인간의 직관에는 잘못 레이블링된 데이터가 있을 수 있다. 그래서 우리는 주어진 행렬과 레이블이 있는 다중 레이블 CNN 분류기를 사용했다. 이를 통하여 레이블을 다시 Self-Ensemble하면서 노이즈가 많을 수 있는 레이블링 방법을 보완하고자 하였다. 각 라벨에 해당하는 이미지의 수가 전부 같은 상황이 아니기 때문에 가중치 밸런싱 기술을 적용하여 학습을 수행하였다. 한편, 각 마디 행렬은 이 분류기에 의해 재분류되어 각 마디 행렬이 가장 적합한 기술 레이블을 갖도록 데이터 세트를 구성했다. 그런 다음 이 데이터를 FLAGNet의 생성모델을 학습시키는 데에 사용한다.

음악 기술 레이블	조건
repeating	전체 음표 중 $n\%$ 이상의 음표의 음고가 동일할 때 repeating으로 정의한다.
up_stepping	전체 음표의 $n\%$ 이상이 음고 변화 2 이하로 상승할 때 up_stepping으로 정의한다.
down_stepping	전체 음표의 $n\%$ 이상이 음고 변화 2 이하로 하락할 때 down_stepping으로 정의한다.
up_leaping	전체 음표의 $n\%$ 이상이 음고 변화 3 이상으로 상승할 때 up_leaping으로 정의한다.
down_leaping	전체 음표의 $n\%$ 이상이 음고 변화 3 이상으로 하락할 때 down_leaping으로 정의한다.
stepping_twisting	전체 음표의 $n\%$ 이상이 음고 변화 2 이하로 오르고 내리는 형태일 때 stepping_twisting으로 정의한다.
leaping_twisting	전체 음표의 $n\%$ 이상이 음고 변화 3 이상으로 오르고 내리는 형태일 때 leaping_twisting으로 정의한다.
fast_rhythm	1마디 안에 음이 9개 이상일 경우 fast_rhythm으로 정의한다.
one_rhythm	전체 음표의 실 연주 시간, 즉 다음 음표까지의 시간이 동일하다면 one_rhythm으로 정의한다.
triplet	셋잇단음이 포함되어 있으면 triplet으로 정의된다.
staccato	전체 음표의 $n\%$ 이상의 연주시간이 최소시간(1/6 timestep)인 경우 staccato로 정의한다.
continuing_rhythm	마디에 휴식 시간이 없으면 continuing_rhythm으로 정의한다.
no_skills	위에 것에 해당되는 것이 없거나, 음표의 수가 3개 이하일 경우 'no_skills'라고 정의하며 생성모델에서 사용하지 않는다.

표 4.1: 음악 기술 레이블 알고리즘

4.3 상대적 음고 기반 피아노 롤 인코딩에 관한 추가 실험

상대적 음고 기반 피아노 롤 인코딩 방식이 피아노 롤 기반 인코딩 방식과 비교하여 어떤 성능을 지니는지 검증하기 위해 3가지 추가 실험을 진행해보았다.

첫 째는 분류기 모델이다. 이는 위에서 설명된 Pseudo-레이블을 다시 Self-Ensemble하는 분류기와 동일한 분류기 모델이며, 상대적 음고 기반 피아노 롤 인코딩을 사용한 이미지와 일반적인 피아노 롤 기반 인코딩을 사용한 데이터 모두로 학습을 진행시켜서 성능을 검증해보았다. 두 번째는 클러스터링 모델이다. 우리는 상대적 음고 기반 피아노 롤 인코딩 방식으로 처리된 데이터들을 모아 딥 클러스터링 모델을 학습시켰다. [30]. 클러스터 수는 5로 하였다.

마지막으로 상대적 음고 기반 피아노 롤 인코딩과 일반적인 피아노 롤 인코딩을 한 데이터를 사용하여 VQ-VAE [31] 모델을 학습시켰고, 둘의 성능을 비교하였다. 두 모델 모두 임베딩의 갯수는 64개로 학습을 하였다.

4.4 FLAGNet

Multi-vector Conditional Deep Convolutional GAN을 학습할 때, 이 모델의 성능을 향상시키는 몇 가지 방법이 있다. 데이터 스무딩 기술을 적용하고, 생성기와 판별기 간의 학습 균형을 제어함으로써 GAN 성능을 향상시킬 수 있다 [32]. 본 연구의 경우, 생성기와 판별기 간의 균형을 유지하기 위해 판별기의 학습 정도를 생성기의 0.1배 정도로 설정하여 비교적 학습 속도가 느린 생성기와, 빠른 판별기 간의 균형을 조정하였다. 우리는 또한 GAN의 Adversarial Loss와 pix2pix구조의 L1 Condition Loss 사이의 균형을 유지하기 위해 Condition Loss에 곱해지는 λ 값을 0.0005로 사용했다. 자세한 생성기 G와 판별기 D의 모델 구성은 그림4.1, 그림4.2와 같다.

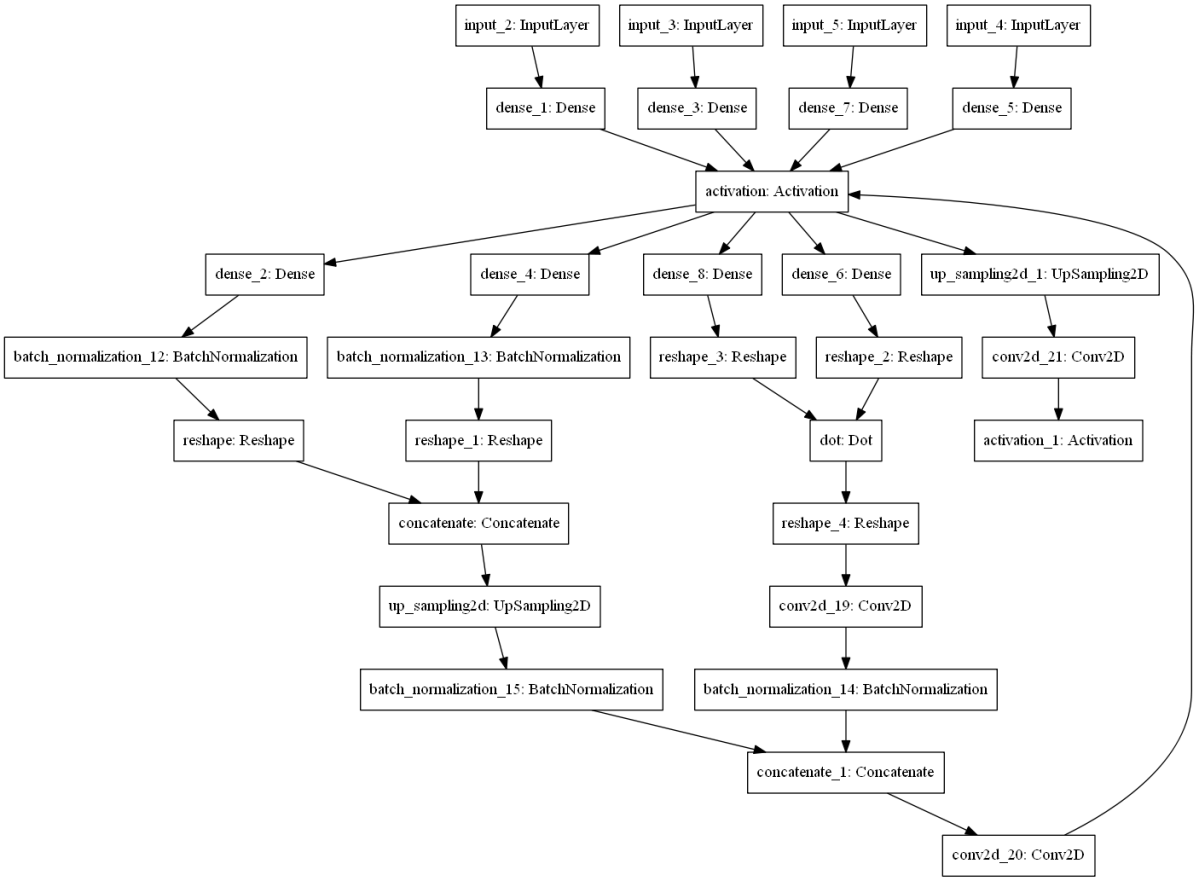


그림 4.1: MCDCGAN의 생성기 구현

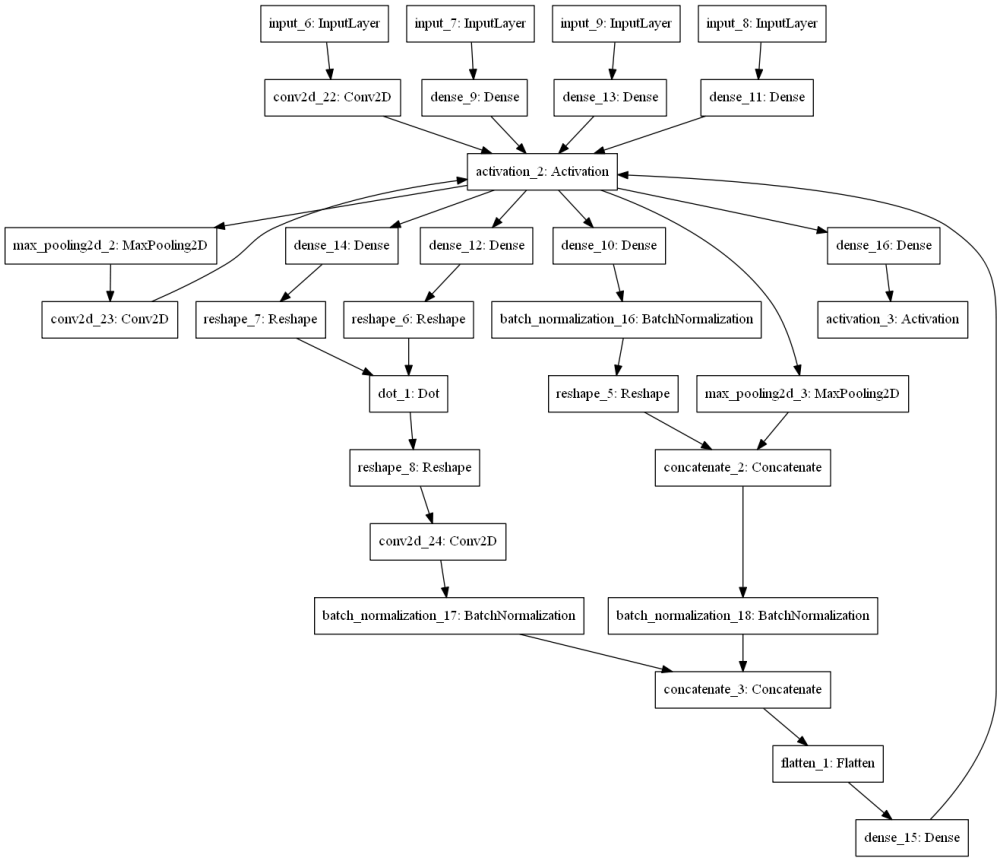


그림 4.2: MCDCGAN의 판별기 구현

기술 레이블 시퀀스 RNN은 LSTM [33] 레이어 및 GRU 레이어 [34]를 통해 성능을 더욱 향상시킨다. 실제 생성 과정에서 다음 스킬을 정의할 때 RNN 모델의 예측은 정규화 되고 이를 확률 분포로 사용하여 다음 스킬을 결정한다. 또한 동일한 스킬이 과도하게 반복되지 않도록 동일한 스킬이 반복될 경우 정규화 시의 예측 값을 $1/3$로 감소시킨다.

디코딩 과정에서 코드 점수를 계산하기 위해 코드 노트를 생성했다. 로마 숫자 분석에 기반한 7가지 디아토닉 코드는 $I, ii, iii, IV, V, vi,\ \ vii^o$이다. [35]우리는 $\ vii^o$를 제외하고 6개의 코드를 사용한다. 이들은 각 마디에 대해 무작위로 선택되며 기본 코드로 사용된다. 이를 통해 12개의 기본 메이저 차트 음계와 각 막대에 대한 코드 음계를 통해 보다 완성도 높은 음악을 만들어 낼 수 있다. 마이너 스케일은 디아토닉 코드를 마이너 스케일로 주어 유사한 방식으로 구현될 수 있다. 이 과정에서 사용자가 직접 코드를 condition으로 줄 수도 있지만, 일반적인 생성 과정에서는 코드에 대하여 무작위 선택을 사용했다.

주어진 바의 모양에 따라 Relational Pitch 피아노 롤에서 다음 바에서 음높이가 올라갈지 내려갈지를 결정하는 것이 필요하다. 우리는 음고 변화 분류기에 간단한 CNN 기반 이진 분류기 모델을 사용한다. 음고 변경 레이블에는 Up, Down, Meaning_less Last_bar의 네 가지 경우가 있다.

Meaning_Less 레이블은 피치 변경이 다음 마디에서 발생하지 않거나 다음 마디가 완전히 음을 연주하지 않는 경우에 사용된다. Last_bar 레이블은 각 곡의 마지막 마디에 사용된다. 이 이진 분류기 모델은 Up 레이블 또는 Down 레이블을 가진 데이터로만 훈련되었다.

이미지 생성 및 디코딩 후 MIDI 처리 과정에서 방법 외에도 음악적 세부 사항에

대한 몇 가지 요소를 구현했다. 코드 스케일에서 벗어나는 음들을 제어하거나, 화음이 생성된 경우에 이를 단선율로 바꾸는 등의 간단한 사용자의 요구를 반영할 수 있도록 하였다.

4.5 결과

이 절에서는 제안한 인코딩 방식인 상대적 음고 기반 피아노 롤 인코딩에 관한 추가 실험의 결과에 대하여 분석하고, 훈련한 모델에 대해 시행한 정성적 평가에 대해 분석한다.

4.5.1 상대적 음고 기반 피아노 롤 인코딩에 관한 Ablation study

우리는 POP909 데이터 셋을 이용하여 피아노 롤 데이터와 Relational Pitch 피아노 롤 데이터에 대한 다중 레이블 분류기를 훈련시켰다. 두 분류기 모두 동일한 구조와 동일한 hyperparameter 하에서 학습을 시켰다.

지표	Relative Pitch 피아노 롤	피아노 롤
Accuracy	0.2966	0.2393
Precision	0.8298	0.7968
Recall	0.7636	0.7345
Time per epoch(s)	11.6	12.8

표 4.2: Validation Data에 대한 데이터 분류 성능

표 4.2는 상대적 음고 피아노 롤 데이터를 사용한 모델 성능이 일반 피아노 롤 데이터를 사용한 경우보다 지표가 뛰어남을 보여준다. 희소 데이터를 촘촘하게 표현할 수 있어 성능이 향상됐고, 학습 시간도 빨라질 수 있는 것으로 보인다.

또한 데이터들을 클러스터링 하는 모델을 학습시킨 후에, 클러스터링 된 결과를 그림 4.3에 t-SNE Visualization [36]하였다.

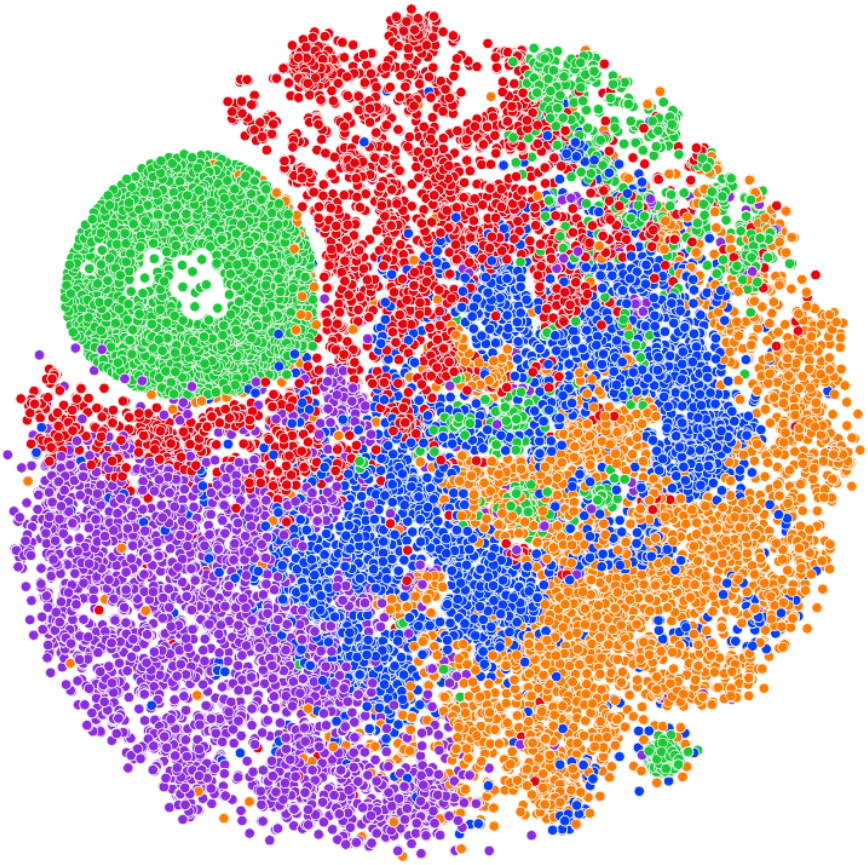


그림 4.3: 클러스터링 결과 t-sne 시각화.

주어진 클러스터 수인 5개에 대하여, 각각 구별이 가능이 될 수 있도록 결과가 나타난 모습이다.

상대적 음고 피아노 롤 데이터와 일반적인 피아노 롤 데이터를 이용하여 VQ-

VAE모델을 학습시켰다. 그 결과 두 경우 모두 Mean Squared Error Reconstruction loss가 0.0011까지 도달하였다. Reconstruction 예시가 그림 4.4에 제시되어 있다.

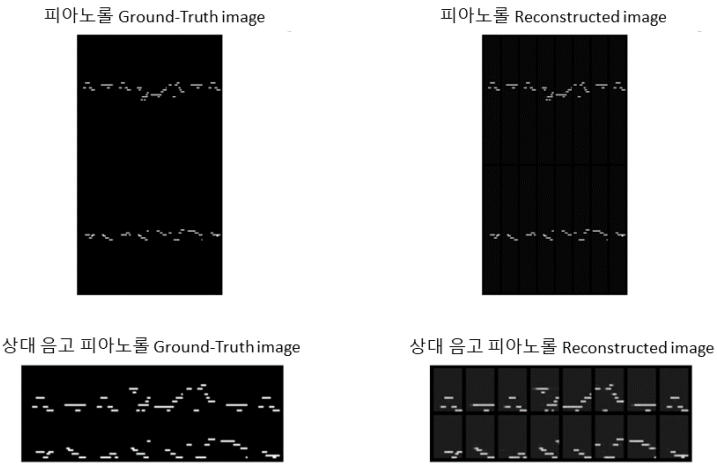


그림 4.4: VQVAE모델 Reconstruction 결과

4.5.2 FLAGNet 작곡 모델 평가

Mturk(Amazon Mechanical Turk) [37]에서 FLAGNet의 자동 음악 생성 성능을 확인하기 위해 30명이 설문조사에 참여[1]하였으며, 실제 설문조사는 FLAGNet 모델로 제작된 4가지 심볼 멜로디, 그 중 2개는 Multi-vector Conditional Deep Convolutional GAN의 결과로부터 선택되었고, 나머지는 생성 모델과 함께 스킬 레이블 조건만 사용한 DCGAN의 결과로부터 선택되었다. 그리고 MidiNet 모델에서 생성된 MIDI 샘플이 두 개를 선택하였고, POP909 데이터 셋의 두 멜로디를 선택하였다. 우리는 001과 002라는 이름의 두 데이터를 선택한다. 원래의 멜로디로, 우리는 다른

[1]IRB 승인 완료, IRB 번호 No. 2206/002-003

생성된 음악 작품들처럼 각 막대에 하나의 코드 음을 가지도록 코드 진행을 수정했다. 모든 곡들은 동일한 가상 피아노 악기를 사용하여 wav 파일로 내보내졌다. 각 노래의 길이는 8마디이다. 참가자들은 어떻게 노래가 만들어졌는지, 누가 작곡했는지 알 수 없지만, 어떤 음악은 AI 모델에 의해 만들어졌고, 어떤 음악은 사람에 의해 만들어졌다는 것만 알고 있다. 이들은 이 곡들을 네 가지 범주에 대해 최대 5점으로 평가한다. 창의성, 편안함, 현실성, 그리고 음악성. 각 범주를 다음과 같이 설명했다.

창의성(Creativity) 주어진 리듬을 활용할 가능성이 크거나 이전에 존재하지 않았다고 생각되면 창의성 점수가 높다.

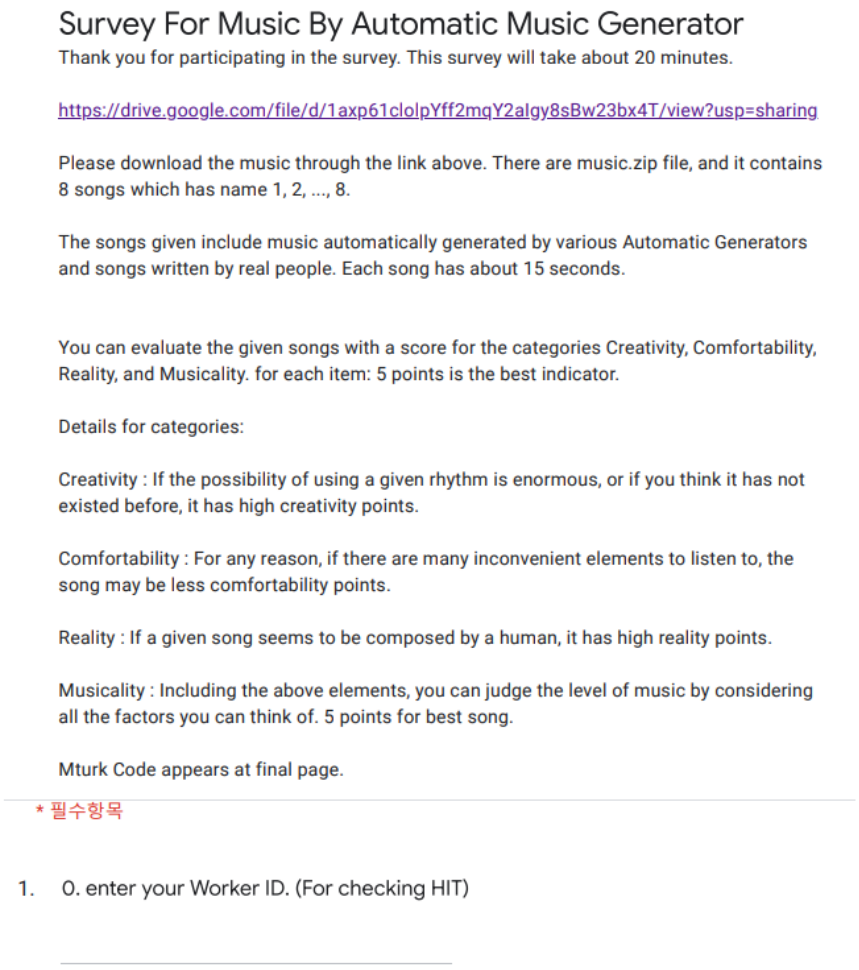
편안함(Comfortability) 어떤 이유로든 듣기 불편한 요소가 많다면 편안함 수치가 낮아진다.

현실성(Reality) 주어진 노래가 사람에 의해 작곡된 것처럼 보인다면, 그것은 높은 현실성 점수를 갖는다.

음악성(Musicality) : 위의 요소를 포함하여 생각할 수 있는 모든 요소를 고려하여 음악의 수준을 판단할 수 있다. 높을수록 좋은 음악이다.

실제로 설문 참여자들은 그림4.5와 같은 화면에서 설문조사를 진행하였으며, 음원을 별도로 다운 받아 직접 들어서 진행할 수 있도록 하였다. 다음은 Mturk 설문조사 결과를 Violin Plot으로 나타낸 결과와, 평균값들을 나타낸 표이다.

FLAGNet의 점수는 모든 지표에서 MIDINet에 비하여 상대적으로 높다. 또한 FLAGNet은 POP909 Dataset의 Melody에 비해서도 충분히 좋은 점수를 받는다. 편안함 지수의 경우 MCDCGAN을 이용한 FLAGNet이 가장 많은 점수를 받았다. 이것은 선제 되는 마디의 음을 고려하는 것이 음악의 편안함을 위해 중요하다는 것을 보여준다. 또한 FLAGNet이 생성한 음악은 창의성, 현실성, 음악성 점수에서 Real


Survey For Music By Automatic Music Generator

Thank you for participating in the survey. This survey will take about 20 minutes.

https://drive.google.com/file/d/1axp61clolpYff2mqY2algy8sBw23bx4T/view?usp=sharing

Please download the music through the link above. There are music.zip file, and it contains 8 songs which has name 1, 2, ..., 8.

The songs given include music automatically generated by various Automatic Generators and songs written by real people. Each song has about 15 seconds.

You can evaluate the given songs with a score for the categories Creativity, Comfortability, Reality, and Musicality. for each item: 5 points is the best indicator.

Details for categories:

Creativity : If the possibility of using a given rhythm is enormous, or if you think it has not existed before, it has high creativity points.

Comfortability : For any reason, if there are many inconvenient elements to listen to, the song may be less comfortability points.

Reality : If a given song seems to be composed by a human, it has high reality points.

Musicality : Including the above elements, you can judge the level of music by considering all the factors you can think of. 5 points for best song.

Mturk Code appears at final page.

* 필수항목

1. 0. enter your Worker ID. (For checking HIT)

그림 4.5: Mturk 설문조사 링크의 화면.

Music과 비슷한 수준의 점수를 얻었다.

전반적으로, FLAGNet은 각 바의 음악적 기술, 그리고 각 바의 코드를 다루면서 음악을 생성할 수 있다. 창의적이고 사실적인 멜로디를 연출할 수 있고 음악성도 뛰

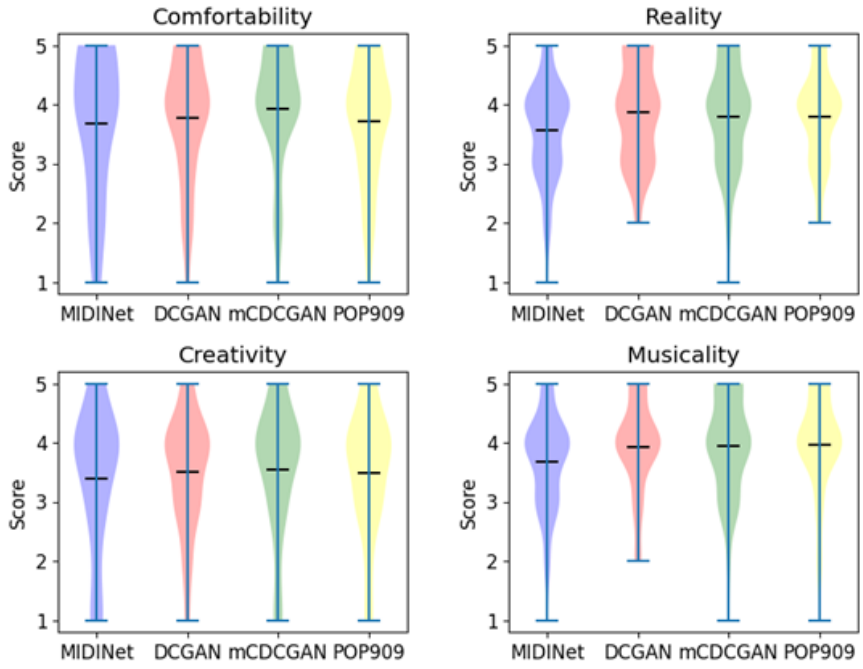


그림 4.6: Mturk 설문조사 결과의 Violin Plot.

어난 것을 확인할 수 있다.

Songs	Creativity	Comfortability	Reality	Musicality
데이터 셋 음악(POP909)	3.483	3.717	3.783	**3.967**
Midinet 음악	3.4	3.683	3.567	3.683
FLAGNet 음악(DCGAN)	3.5	3.767	**3.867**	3.917
FLAGNet 음악(MCDCGAN)	**3.55**	**3.933**	3.8	3.95

표 4.3: Mturk 설문조사 결과의 평균을 정리한 표.

제 5 장 결 론

5.1 연구 의의

우리는 심볼릭 음악을 마디별로 처리하여 음악적인 기술을 고려하여 MIDINet과 같은 기존 생성 모델보다 창의적이고 음악적인 음악을 만드는 데 성공한 음악 생성 모델인 FLAGNet을 제안하였다. 이 모델을 제안함으로써, 우리는 마디 단위의 음악이 가지고 있는 기술의 적용을 통해 완성된 음악이 생성될 수 있음을 보여주었다.

또한 새로운 심볼릭 음악 인코딩 방식인 상대 음고 기반 피아노 롤을 제안하였고, 이가 다양한 모델에 활용될 수 있으며 더 좋은 성능을 보일 수 있음을 보여주었다.

음악 생성과 관련된 영역 외에도, FLAGNet 모델의 접근 방식은 대량의 데이터를 다루는 영역에서 유용할 수 있다고 생각된다. 데이터를 더 작은 조각으로 나누고, 더 작은 단위를 기준으로 분석하고, 이들의 관계를 파악하는 방법에 대한 추가 연구가 기대된다.

위의 연구에서는 마디별로 휴리스틱 알고리즘을 기반으로 스킬의 정의를 수행하고 분류기로 재표시를 하였으나, 보다 정확한 방법으로 라벨링을 하거나 'buildup'이나 'verse'와 같은 음악 구조로 분류하는 등 보다 실용적인 라벨링을 준비하여 성능을 높이거나 활용할 수 있을 것으로 기대된다. 또한 본 연구에서 학습은 단성 리듬과 솔로 트랙 음악에만 적용되었다. 상대 음고 기반 피아노 롤을 사용하여 멀티 트랙 음악을 생성하는 것은 향후 작업으로 해결될 것으로 기대된다. 또한 상대 음고 기반 피아노 롤은 음악 생성 작업뿐만 아니라 장르 분류, 스타일 변환 등 Music Information Reitrival 연구 분야의 심볼릭 음악 영역에서 넓은 범위로 활용될 것으로

기대된다.

5.2 한계점

상대 음고 기반 피아노 롤이 현존하는 피아노 롤의 단점을 전부 상쇄시킬 수는 없다. 특히나 음고가 정수이기 때문에 생기는 문제는 여전히 존재한다. 예시를 들면 C4, C#4, D4 3개의 음이 48, 49, 50이라는 정수로 대응되는 문제로 인해 학습 과정에서 문제가 생기는 것이다. 실제로는 화성의 구조 등을 고려할 수 있어야하나, 이를 단순히 정수로 Representation할 경우 모델이 이를 고려하기 어려워진다. 본 연구에서 활용한 방식을 활용할 경우 전체적인 음고를 높이거나 낮추는 Shifting 과정에서 생기는 문제에는 Robust하지만, 한 마디 내에서 실제 음이 어떻게 변경되는지에 대해서 완전히 Robust하지는 않다.

생성 모델인 FLAGNet 경우 멜로디 생성만을 진행하는 형태로 구성되어있는데, 실제로는 하나의 악기가 다양한 트랙을 연주할 수도 있고 여러가지 악기가 사용될 수도 있다. 하지만 현재 모델의 구조만으로는 이를 해내기 까다롭다. 또한 각 마디에 주어지는 음악 기술 레이블이 휴리스틱 알고리즘으로 주어지기 때문에 레이블 자체가 다소 잘못 주어졌을 가능성을 배제할 수 없다.

5.3 향후 연구

상대 음고 기반 피아노 롤을 활용함에 있어, 절대 음고를 지정해주는 것이 정말 중요한 작업이다. 현재는 음악 이론에 기반하여 알고리즘 기반 작업을 진행하는 형태이지만, 이를 조금 더 개선할 방식을 찾을 수 있을 것이라고 생각된다. 또한 절대 음고를 지정함에 있어 현재는 음고의 높낮이 범위를 지정하는 것이 이전 마디의 방향 분류기 결과에 따라 정해지게 되는데, 실제로 활용하는 과정에서 이전 마디 자체

가 존재하지 않을 수도 있으며, 현재의 음고의 높낮이 범위를 지정하는 알고리즘이 안정적인 음악을 음고의 범위를 줄 수 있다는 보장이 없어 이를 개선하는 연구도 할 수 있을 것으로 보인다.

또한 음악을 생성하는 FLAGNet에 대해서는 다양한 악기를 다루는 Multitrack에서의 Task를 진행할 수 있을 것으로 보인다.

음악을 생성하는 연구 뿐만이 아니라 다른 심볼릭 음악을 활용하는 연구에 상대음고 피아노 롤을 적용할 수 있으며, 또한 주어진 음악 데이터를 마디별로 나누어서 분석하는 아이디어를 다양한 연구에 사용할 수 있을 것으로 생각한다.

참고 문헌

[1] R. A. Moog, "Midi: musical instrument digital interface," *Journal of the Audio Engineering Society*, vol. 34, pp. 394–404, 1986.

[2] M. Dua, R. Yadav, D. Mamgai, and S. Brodiya, "An improved rnn-lstm based novel approach for sheet music generation," *Procedia Computer Science*, vol. 171, pp. 465–474, 2020.

[3] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," *arXiv preprint arXiv:1709.06298*, 2017.

[4] N. Kotecha, "Bach2bach: Generating music using a deep reinforcement learning approach," *arXiv preprint arXiv:1812.01060*, 2018.

[5] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," *arXiv preprint arXiv:1809.04281*, 2018.

[6] T. Akama, "Controlling symbolic music generation based on concept learning from domain knowledge.," pp. 816–823, 2019.

[7] D. von Rütte, L. Biggio, Y. Kilcher, and T. Hoffman, "Figaro: Generating symbolic music with fine-grained artistic control," *arXiv preprint arXiv:2201.10936*, 2022.

[8] H. H. Tan and D. Herremans, "Music fadernets: Controllable music generation based on high-level features via low-level feature modelling," 7 2020.

[9] J. Wu, C. Hu, Y. Wang, X. Hu, and J. Zhu, "A hierarchical recurrent neural network for symbolic melody generation," *IEEE Transactions on Cybernetics*, vol. 50, pp. 2749–2757, 2019.

[10] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, "Midinet: A convolutional generative adversarial network for symbolic-domain music generation," *arXiv preprint arXiv:1703.10847*, 2017.

[11] S. Tanberk and D. B. Tükel, "Style-specific turkish pop music composition with cnn and lstm network," pp. 181–185, 2021.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," pp. 1125–1134, 2017.

[14] S. Li, S. Jang, and Y. Sung, "Melody extraction and encoding method for generating healthcare music automatically," *Electronics*, vol. 8, p. 1250, 2019.

[15] J. Wu, C. Hu, Y. Wang, X. Hu, and J. Zhu, "A hierarchical recurrent neural network for symbolic melody generation," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2749–2757, 2019.

[16] M. Dua, R. Yadav, D. Mamgai, and S. Brodiya, "An improved rnn-lstm based novel approach for sheet music generation," *Procedia Computer Science*, vol. 171, pp. 465–474, 2020.

[17] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1180–1188, 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] S. Han, H. Ihm, M. Lee, and W. Lim, "Symbolic music loop generation with neural discrete representations," *arXiv preprint arXiv:2208.05605*, 2022.

[20] S. Li, S. Jang, and Y. Sung, "Melody extraction and encoding method for generating healthcare music automatically," *Electronics*, vol. 8, no. 11, p. 1250, 2019.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[23] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[24] S.-L. Wu and Y.-H. Yang, "Musemorphose: Full-song and fine-grained music style transfer with one transformer vae," *arXiv preprint arXiv:2105.04090*, 2021.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," pp. 2672–2680, 2014.

[26] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "Self: Learning to filter noisy labels with self-ensembling," 10 2019.

[27] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, X. Gu, and G. Xia, "Pop909: A pop-song dataset for music arrangement generation," *arXiv preprint arXiv:2008.07142*, 2020.

[28] J. S. Downie, "The music information retrieval evaluation exchange (mirex)," *D-Lib Magazine*, vol. 12, 2006.

[29] C. Raffel, *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. PhD thesis, Columbia University, 2016.

[30] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoencoders," in *International conference on neural information processing*, pp. 373–382, Springer, 2017.

[31] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[32] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," pp. 2234–2242, 2016.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[34] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[35] J. Mehegan, *Tonal and Rhythmic Principales: Jazz Improvisation I*. Watson-Guptill Publications, 1984.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[37] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?," 2016.

ABSTRACT

The technology for automatic music generation has been very actively studied in recent years. However, almost in these studies, handling domain knowledge of music was omitted or considered a difficult task. In particular, research that analyzes and applies the characteristics of each bar of music is rare, even though it is essential in human composition. We propose a model that generates music by handling the musical characteristics of bars and priming note conditions. We first analyze symbolic music data as piano-roll based method with a relational pitch approach, which increases the utilization of the piano-roll based MIDI encoding method and enables the use of generational results extensively. We have trained a model to generate these data with priming notes condition and musical skill label, by the multi-vector conditional deep convolutional generative adversarial network. The part related to the musical skill condition, we analyzed the good combination of the sequence of which characterized bars, simply done by Recurrent Neural Network with Long short Term Memory and Gated Recurrent Unit layer. While handling inputs like a minimum unit of note, length of music, or chart scales, the resulting model FLAGNet can generate impressive symbolic music.

Keywords: Music Generation Model, Symbolic Music, Relational Pitch Pianoroll Encoding, Multi-vector Conditional Deep Convolutional Generative adversarial Network

Student Number: 2021-24957

감사의 글

처음 본 연구를 진행하고, 글로 옮겼을 때에 부족한 점이 정말 많았습니다. 이를 보완해나가는 과정에 MARG 연구원 분들과 이교구 교수님의 지도가 정말 큰 도움이 되어, 감사의 말씀을 전해드리고 싶습니다.
현재도 많이 부족하다고 생각됩니다. 하지만 MARG 연구실에서 지내면서 앞으로 하고 싶은 연구를 어떤 것을 할지, 어떻게 연구를 할지, 또한 그 연구를 어떻게 다른 사람들에게 소개할 지에 대해서 방향을 잡을 수 있었던 것 같습니다. 다시 한 번 교수님과 연구원분들께 감사의 말을 전합니다.