Symbolic Music Loop Generation with Neural Discrete Representations

Sangjun Han[1], Hyeongrae Ihm[1], Moontae Lee[1,2], and Woohyung Lim[1*]
[1]LG AI Research, Seoul, South Korea,
[2]University of Illinois at Chicago, Chicago, Illinois, USA

Introduction

Loop Generation

- Loops in music are essential ingredient for creating remixes or mash-ups
  - Generating long sequences by utilizing the expressive power of Transformer has limitations, derived from the error accumulation and **rhythmic irregularity**
  - Previous works have attempted to extract loops explicitly, requiring musical feature extraction and heuristic process to decide which features should be more weighted

Discrete Representations for Music

- Musical Ideas can be formed in combinations of **finite symbols**
  - Learning discrete codes is sufficient to represent the continuous world such as objects in vision, words in language, and phonemes in speech
  - Consecutive notes can be also represented as discrete symbol

Contributions

- We propose the framework of symbolic music loop generation, which involves loop extraction, loop generation, and its evaluation
- For **loop extraction**, we design **a structure-aware loop detector** trained by external audio sources to extract loops of 8 bars from MIDI
- For **loop generation**, we verify that **an autoregressive model combined with discrete representations** can generate plausible loop phrases which can be repeated
- With randomly initialized networks for embedding, we evaluate sample quality in terms of fidelity and diversity

Methods

Data Preparation

- Audio loop dataset from Looperman ($x_{wav} \in \mathcal{R}$, 1,000 loops of 8 bars)
- Lakh MIDI dataset, ($x_{MIDI} \in \{0,1\}^{time \times pitch}$, 5,687,274 phrases of 8 bars)

Loop Extraction

- We transform each the x_{wav} and x_{MIDI} to two bar-to-bar correlation matrix indicated as C_{wav} and C_{MIDI} (Fig. 1)
- We train a loop detector using Deep SVDD [1], treating C_{wav} as normal samples (training set) and measure the likelihood of C_{MIDI} (test set)
- At inference time, the loop score of C_{MIDI} can be estimated by the Deep SVDD's distance metric
- We collect 751,935 x_{MIDI} for loops using our loop detector

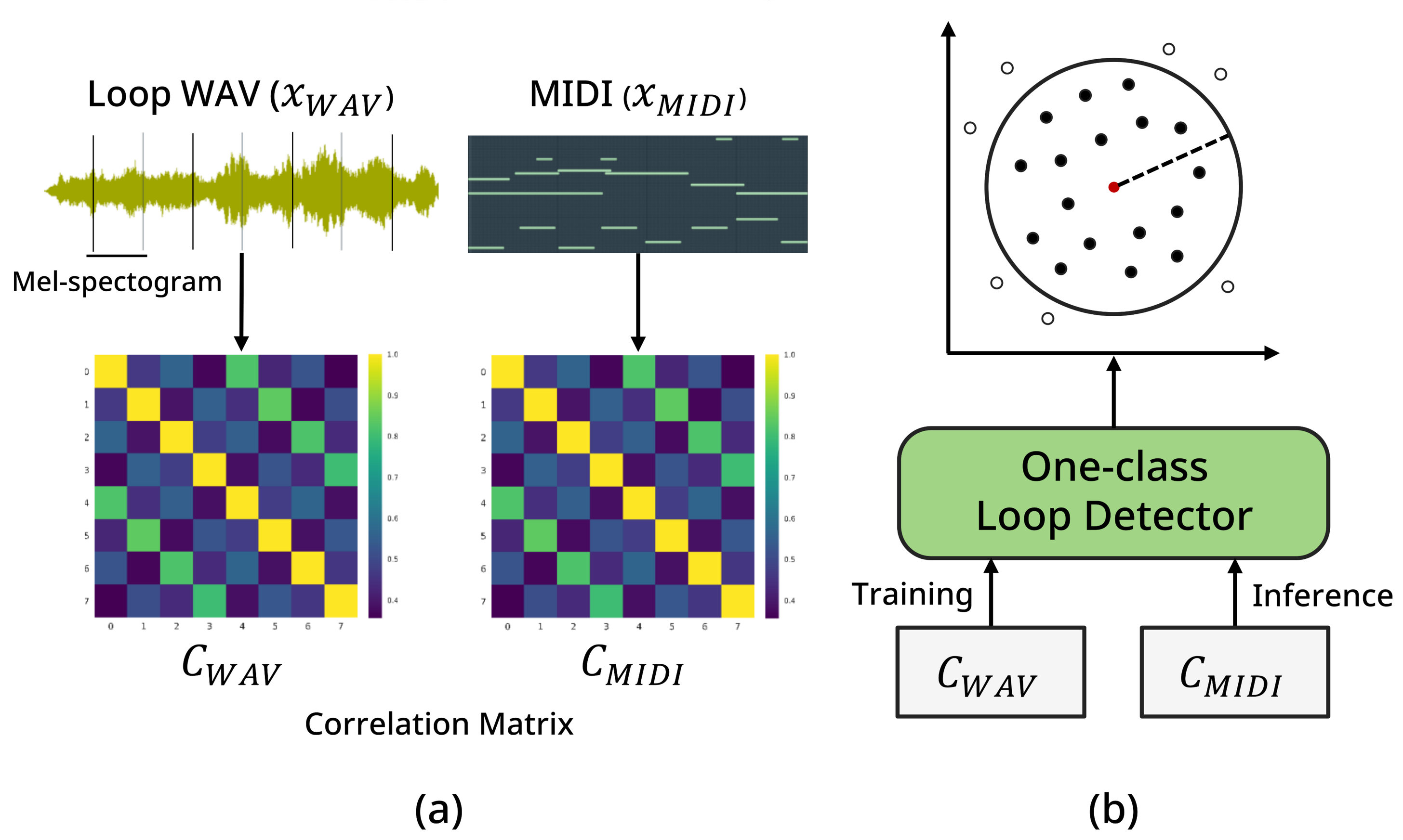

Fig. 1. The process of loop extraction

Loop Generation

- VQ-VAE maps a data sequence into discrete latent space and reconstructs it to the original data space (Fig. 2) [2]
  - Due to finite latent codes, our VQ-VAE is a little worse than CNN-VAE for the reconstruction task
- We design a LSTM autoregressive model over quantized embeddings to generate unseen samples (Accuracy: 76.65 %)

Quantitative Evaluation

- Model metric: related to evaluating the capacity of generative models
  - Reconstruction Error
- Musical style: related to measuring how much our intended properties of the loop are involved in generated samples
  - Loop Score (LS) from our loop detector, Unique Pitch (UP) for harmony, Note Density (ND) for rhythm
- Similarity metrics: related to measuring similarity of true and generated samples on random latent space, indicating the sample fidelity and diversity
  - Precision & Recall (P & R) [3]
  - Density & Coverage (D & C) [4]

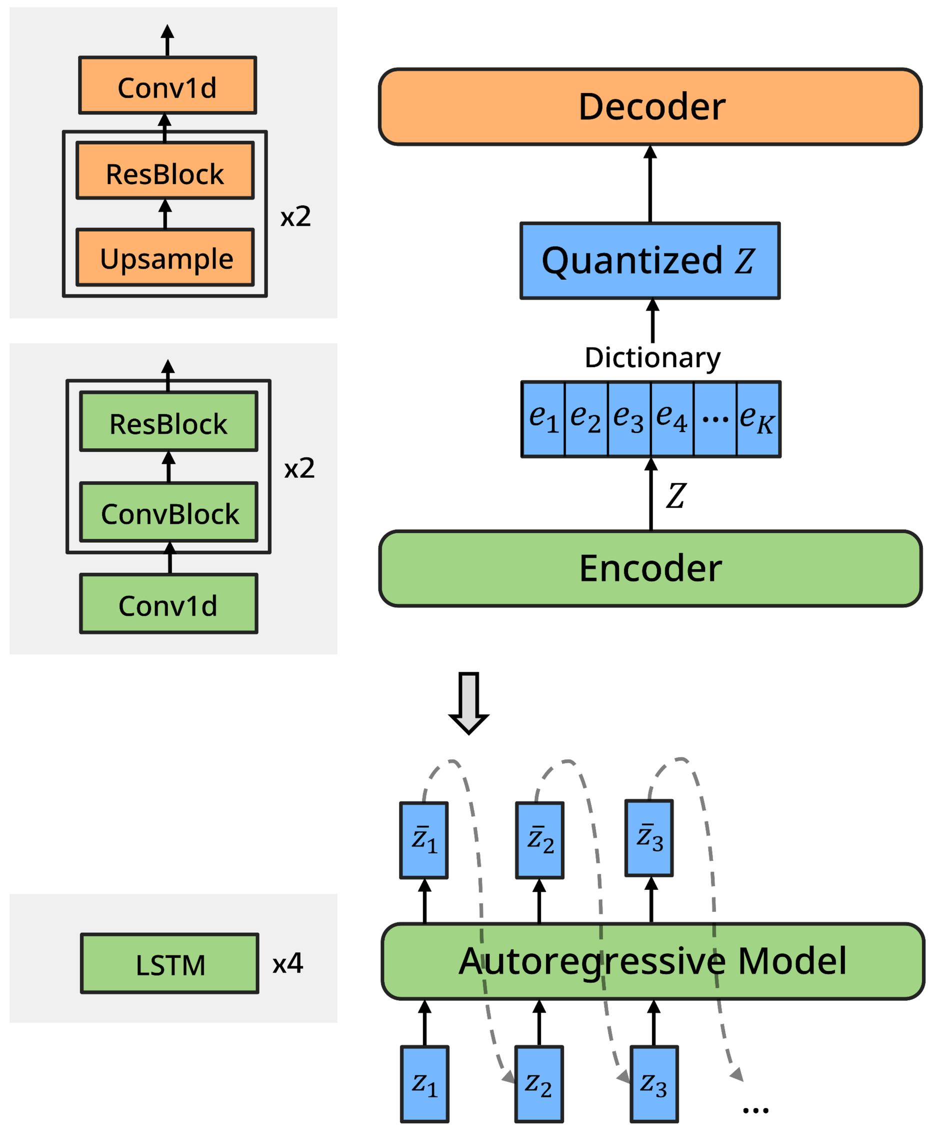

Fig. 2. The process of loop generation

Results

Model	LS	UP	ND	P	R	D	C
Training Set	6.806e-3	5.769	14.383	-	-	-	-
CNN-VAE	3.496e-1	5.614	12.648	0.642±0.033	0.625±0.021	0.617±0.076	0.746 ±0.024
Music Transformer	7.200e-1	4.127	11.290	0.546±0.077	0.359±0.113	0.687±0.174	0.408±0.064
MuseGAN	2.307e-1	**5.790**	14.011	0.641±0.013	**0.689±0.012**	0.673±0.045	0.842±0.013
VQ-VAE+LSTM (temperature sampling)	2.275e-1	5.079	14.289	0.768±0.013	0.655±0.022	1.263±0.047	0.949±0.002
VQ-VAE+LSTM (top-k sampling=30)	**1.978e-1**	5.044	14.320	0.779±0.015	0.636±0.015	1.328±0.072	**0.952±0.004**
VQ-VAE+LSTM (top-p sampling=0.08)	2.037e-1	5.042	**14.341**	**0.783±0.017**	0.638±0.029	**1.337±0.075**	0.950±0.005

- Our proposed model has achieved the highest performance in terms of LS, ND, P, D, and C
- Depending on sampling methods and their parameters, we can observe that there is a trade-off between fidelity and diversity
- The human listening test for 20 people shows that our proposed model can generate musical samples comparable with its training set (Fig. 3)

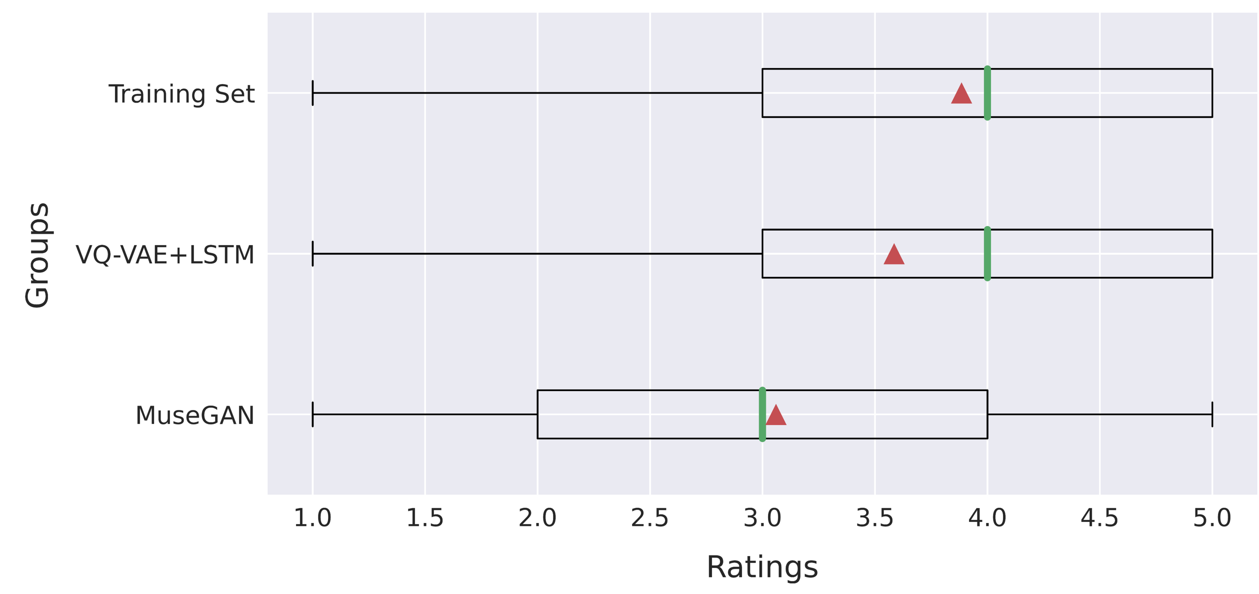

Fig. 3. Human listening test

- Additionally, we verify that the trained loop detector can be used to control the trade-off between fidelity and diversity of the generative models, as known as rejection sampling

Discussion

The Framework of Symbolic Music Generation

- We leverage recurring nature of music by adopting the concept of the loop
- We address two processes; loop extraction and loop generation
- We evaluate our generative models on measuring fidelity and diversity on random latent space

Limitation & Future work

- 1) Generating multi-track instrument, 2) Imposing recurring nature

References

1. Ruff et al., 2018
2. Oord et al., 2017
3. Sajjadi et al., 2018
4. Naeem et al., 2020