MuseGAN: Demonstration of a Convolutional GAN Based Model for Generating Multi-track Piano-rolls

Hao-Wen Dong*, Wen-Yi Hsiao*, Li-Chia Yang, Yi-Hsuan Yang

Music and Audio Computing (MAC) Lab, Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan

`salu133445@citi.sinica.edu.tw, s105062581@m105.nthu.edu.tw, {richard40148, yang}@citi.sinica.edu.tw`

* These authors contributed equally to this work

Introduction

Challenges for music generation:

- **Temporal dynamics**: music is an art of time with a hierarchical structure
- **Multi-track**: each track (instrument) has its own temporal dynamics but collectively they unfold over time in an interdependent way
- **Discrete valued**: it's a sequence of events, not continuous values

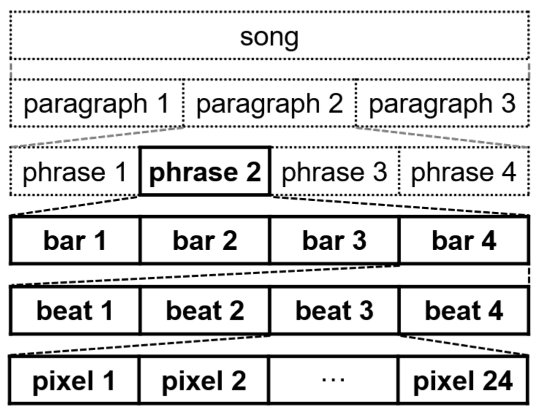


Figure 1. Hierarchical temporal structure of music

MuseGAN (multi-track sequential generative adversarial network) [1] aims to address these 3 challenges altogether. Key points:

- Use **GAN** (specifically WGAN-GP [2]) to support both "conditional generation" (e.g. following a prime melody) and "generating from scratch", following our previous MidiNet model [3]
- Use **convolutions** (instead of RNNs) for speed
- Use a **bar** (instead of a note) as the basic unit for generation
- Learn from **MIDIs** (piano-rolls), not lead sheets
- Experiment with a few network designs for the temporal model and for inter- and intra-track modeling

Demo webpage: `https://salu133445.github.io/musegan/`

Data

The *matched* subset of the Lakh MIDI dataset [4], after cleansing

- Pop/rock, 4/4 time signature, C key
- Five tracks: bass, drums, guitar, piano, strings (others)
- Get 4-bar phrases by structural feature-based segmentation

We are happy to share the data and utility code (go to demo page)!

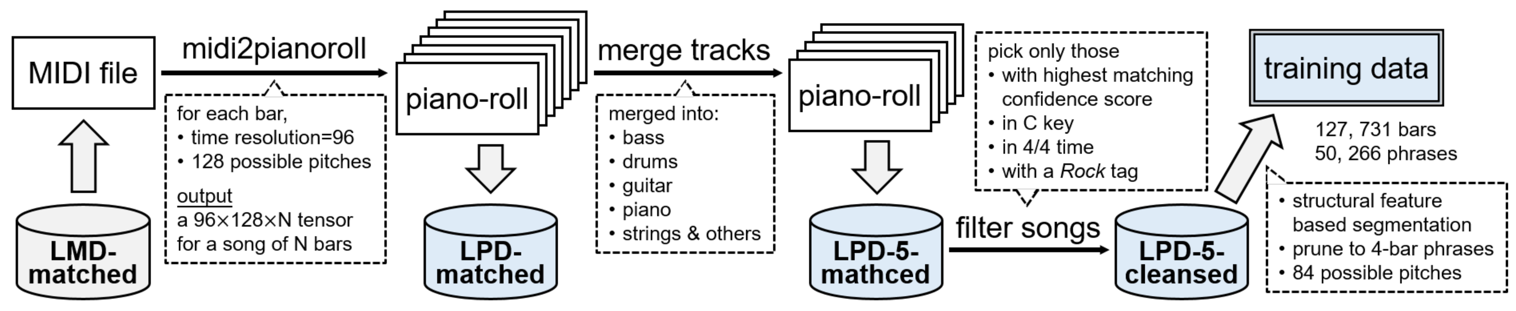


Figure 2. Flowchart of the data cleansing and preprocessing procedure

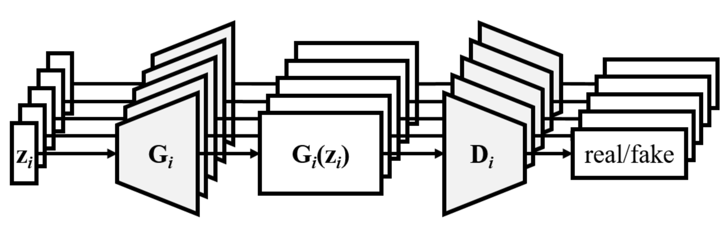
Proposed Model

Modeling the Multi-track Interdependency

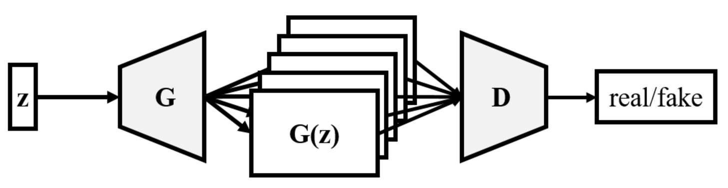
Jamming: Each track has its own generator and discriminator, without any coordination

Composer: All the tracks are generated by one single generator, and critic is given by one discriminator, like a composer or a band leader who evaluate the joint performance of all the musicians (tracks)

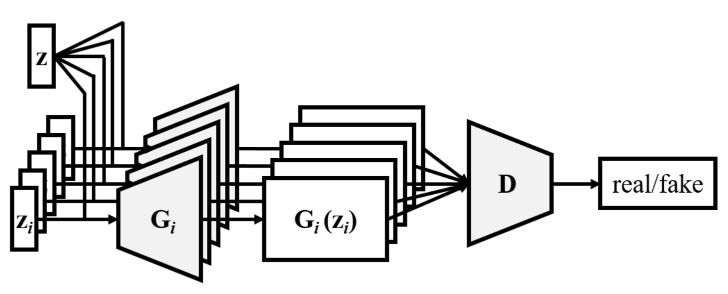
Hybrid: Each track is generated independently by its own generator which takes a shared *inter-track* random vector and a private *intra-track* random vector as inputs; the result is evaluated by one single discriminator



(a) the jamming model



(b) the composer model



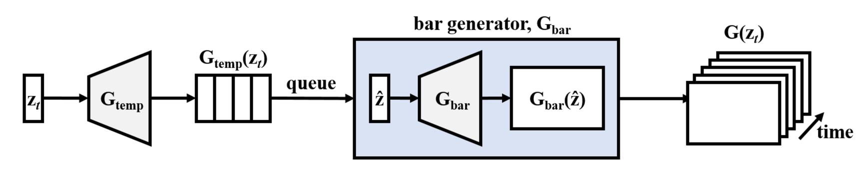
(c) the hybrid model

Figure 3. Multi-track models

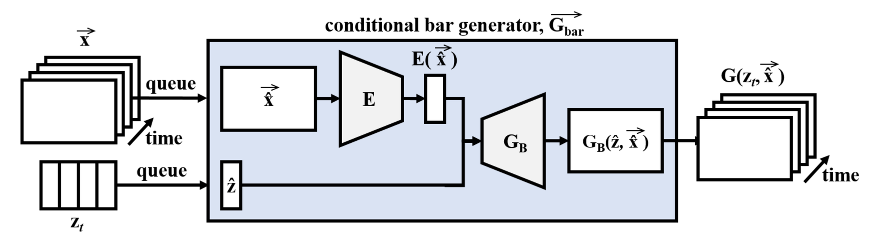
Modeling the Temporal Structure

Generation from scratch: Fixed-length phrases are generated by viewing time as an additional dimension to be generated

Track-conditional generation: by learning to follow the temporal structure of a track given *a priori*



(a) generation from scratch



(b) track-conditional generation

Figure 4. Temporal models

MuseGAN = Temporal models + Multi-track models

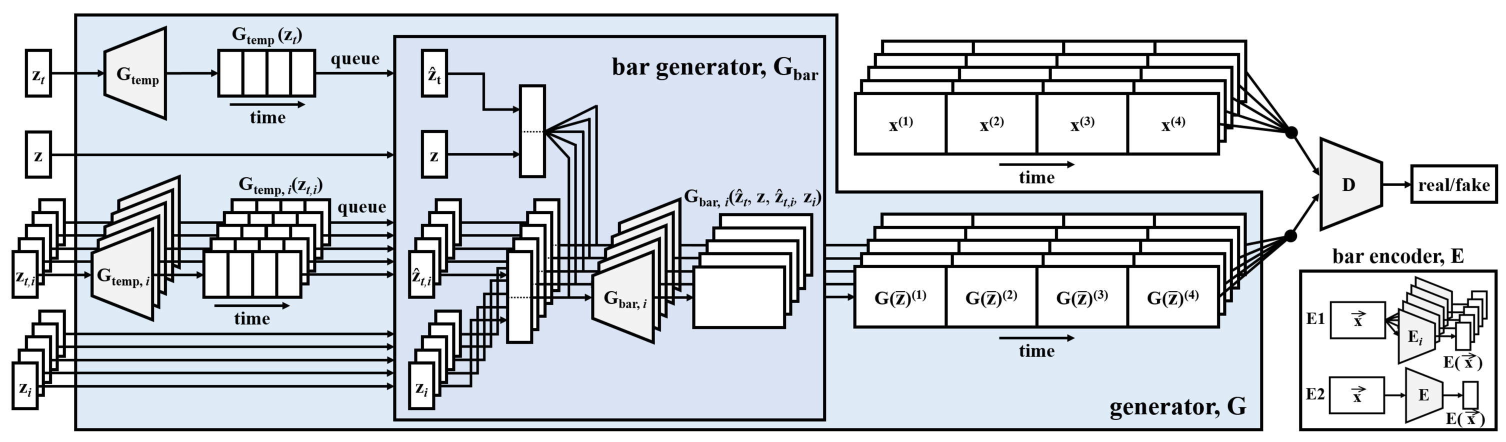


Figure 5. System diagram of the proposed MuseGAN model

Results

1) Sample results (generating from scratch; not cherry-picked):

- The bass is mostly monophonic and playing the lowest pitches
- The drums often have 8- or 16-beat rhythmic patterns
- The other 3 tracks tend to play the chords, and their pitches sometimes overlap (black lines), indicating harmonic relations

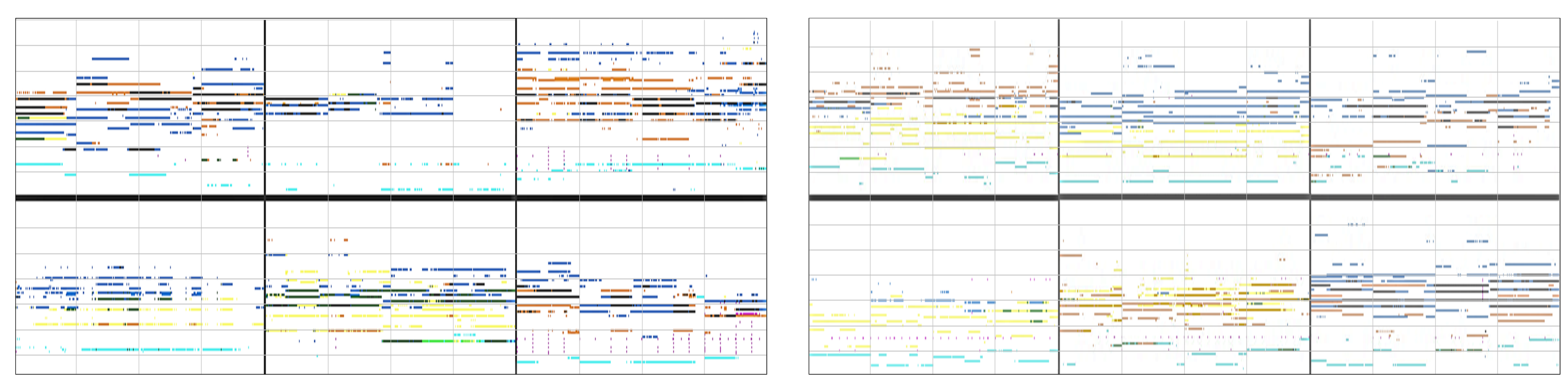


Figure 6. Example generated phrases, left: composer model, right: hybrid model—*cyan*: bass, *purple*: drums, *yellow*: guitar, *blue*: strings, *orange*: piano.

2) The generator becomes better along with the training process:

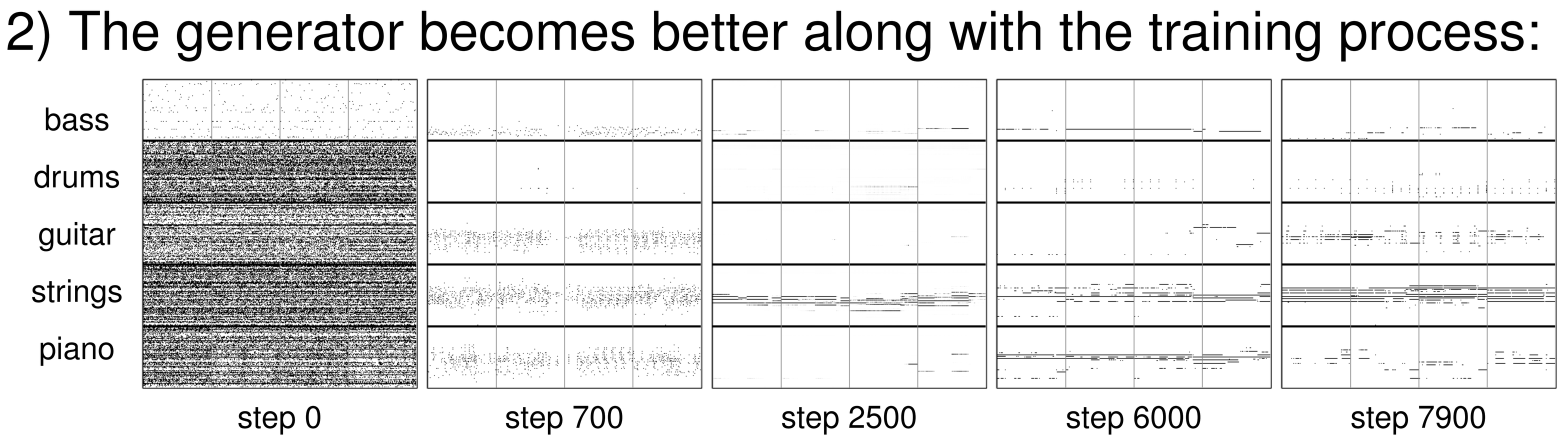


Figure 7. Evolution of a generated phrase (the composer model, from scratch)

Conclusions

- A new convolutional GAN model is proposed for creating binary-valued multi-track sequences; we use it to generate piano-rolls of pop/rock music by learning from a large set of MIDIs
- Still room for improvement so let's further work on it!

References

[1] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. MuseGAN: Symbolic-domain music generation and accompaniment with multi-track sequential generative adversarial networks. *arXiv preprint arXiv:1709.06298*, 2017.

[2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

[3] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. In *ISMIR*, 2017.

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