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
MAC Lab, CITI, Academia Sinica, Taiwan
salu133445@citi.sinica.edu.tw, s105062581@m105.nthu.edu.tw, {richard40148,yang}@citi.sinica.edu.tw

ABSTRACT

Generating realistic and aesthetic pieces is one of the most exciting tasks in the field. We present in this demo paper a new neural music generation model we recently proposed, called MuseGAN. We exploit the potential of applying generative adversarial networks (GANs) to generate multi-track pop/rock music of four bars, using convolutions in both the generators and the discriminators. Moreover, we propose an efficient approach for pre-processing symbolic data and share the data with the community. Our model can generate music either from scratch, or by following (accompanying) a track given by user.

1. INTRODUCTION

As reviewed by Briot *et al.* [1], an increasing number of neural networks have been proposed lately for music generation. However, the task remains challenging due to the following reasons. First, music is an art of time, necessitating a temporal model. Second, music is usually composed of multiple instruments/tracks, with close interaction with one another. Each track has its own temporal dynamics, but they unfold over time interdependently. Lastly, for symbolic domain music generation, the targeted output is sequences of discrete musical events, not continuous values. In our recent work [2], we present an attempt to deal with these issues altogether based on the Wasserstein generative adversarial networks with gradient penalty [3], using convolutions. We present in this demo paper the key ideas of this 'MuseGAN' model (we call it 'MidiNet v2' internally), and refer readers to our arXiv paper for details. [1]

2. DATA

While some existing work used lead sheets or music in ABC format as the training data [1], we aim at learning directly from MIDI files. Specifically, we use a subset of the Lakh MIDI dataset (LMD) [4]. Learning from such a large yet noisy data requires data cleansing. We pick songs that are tagged as 'Rock' songs and discard songs that are not in C key or do not use four-beat time signatures. All MIDI files are converted into the so-called *piano-roll* representations. We categorize all MIDI tracks into five instrument families: bass, drum, guitar, piano and strings and merge the tracks within each instrument family by summing (logical or operation) the corresponding piano-rolls, resulting in piano-rolls of five tracks. For each bar, we set the width (time resolution) to 96 for modeling common temporal patterns such as triplets and 16th notes. We set the height to 84 to cover pitches from `C1` to `C8`. The size of a data tensor per bar is hence 96 (time step) $\times$ 84 (note) $\times$ 5 (track). We consider four bars as a phrase, and accordingly prune longer segments into proper size. The final dataset has in total 127,731 bars, and 50,266 phrases. The goal of our model is to generate a five-track piano-roll of four bars. To facilitate research along this line, we share this cleaned set with the community at: `https://salu133445.github.io/musegan/`.

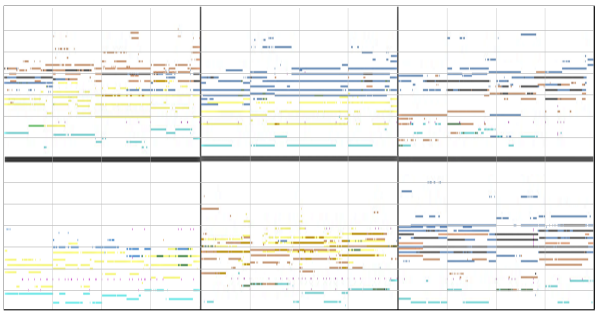

Figure 1: Example result of MuseGAN (hybrid model, from scratch); best viewed in color—**cyan**: bass, **purple**: drum, **yellow**: guitar, **blue**: strings, **orange**: piano.

3. PROPOSED MODEL

3.1 Modeling the Multi-track Interdependency

- **jamming model**: Each track is generated independently by its own generator with its own random input, and critic is given by its own discriminator, just like a teacher specialized in one specific instrument.

[1] HW and WY have equal contribution to this work; LC is currently affiliated with Georgia Tech Center for Music Technology.

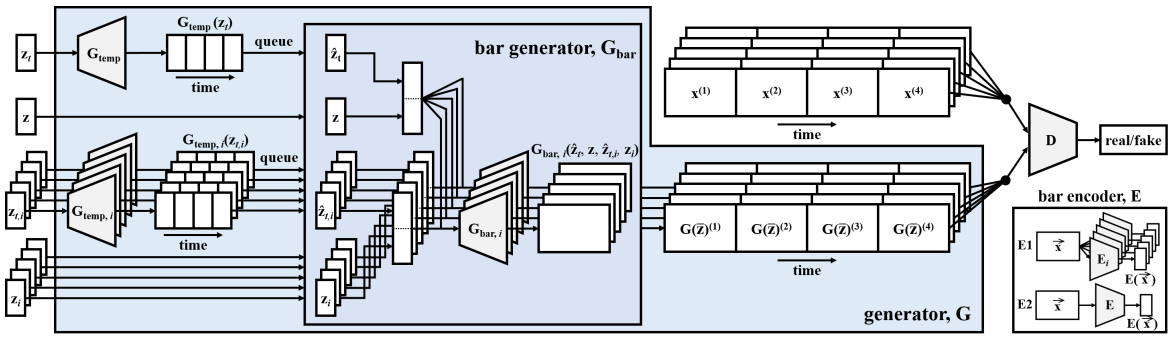

Figure 2: System diagram of the proposed MuseGAN model for multi-track sequential data generation.

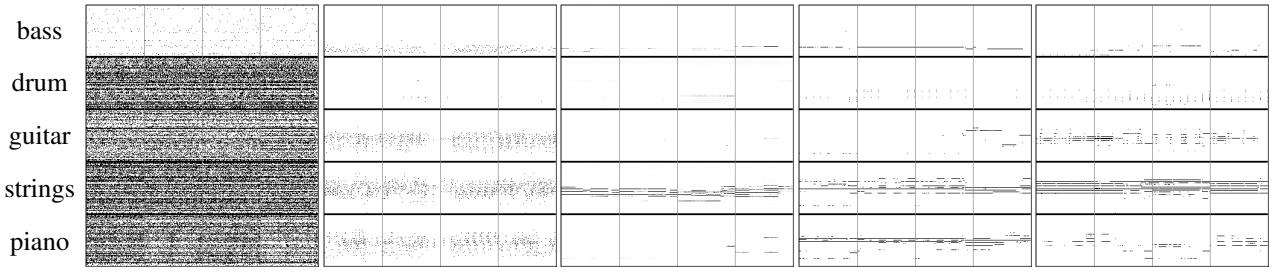

Figure 3: Evolution of the generated piano-rolls (the composer model, from scratch) as a function of update steps.

- **composer model**: All the tracks are generated by a single generator, and critic is given by its one discriminator, which serves as a composer or a band leader and tries to evaluate the joint performance of all the musicians (tracks) in the play.
- **hybrid model**: Each track is generated by its own generator taking a (shared) *inter-track* random vector and a (private) *intra-track* random vector as inputs. This design can offer better controllability. For instance, we can customize the structure for each track without losing any inter-track interdependency.

3.2 Modeling the Temporal Structure

- **generation from scratch**: Fixed-length musical phrases are generated by viewing time as an additional dimension to be generated by the networks.
- **track-conditional generation**: Fixed-length musical phrases are generated by learning to follow the temporal structure of a track given *a priori* by human. This can be applied to human-AI cooperative music generation, or music accompaniment

Incorporating the temporal models with the multi-track models leads to MuseGAN [2], as illustrated in Figure 2.

4. RESULT

Figure 1 shows the piano-rolls of six phrases generated by the hybrid model. We can see that the bass usually plays the lowest pitches and it is mostly monophonic (i.e. playing the melody). The drum often has clear 8- or 16-beat rhythmic patterns. Moreover, the other three tracks tend to play the chords, and their pitches sometimes overlap (creating the black lines), indicating nice harmonic relations. More examples of the generated piano-rolls will be demonstrated on-site and can be found online as well `https://salu133445.github.io/musegan/`.

5. CONCLUSION

We presented a new convolutional GAN model for generating binary-valued multi-track sequences. We have also implemented such a model for generating piano-rolls of pop/rock music by learning from a large corpus of MIDIs. Our model can start to learn things about music, as shown in Figures 1 and 3, but there is still room for improvement. We hope it can inspire more researches along this line.

6. REFERENCES

[1] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. Deep learning techniques for music generation: A survey. *arXiv:1709.01620*, 2017.

[2] 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:1709.06298*, 2017.

[3] Ishaan Gulrajani et al. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

[4] Colin Raffel. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. PhD thesis, Columbia University, 2016. [Online] http://colinraffel.com/projects/lmd/.