1-010

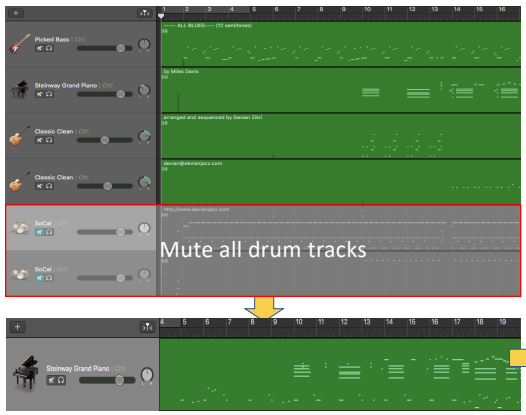
An improvement on CycleGAN based symbolic music genre transfer

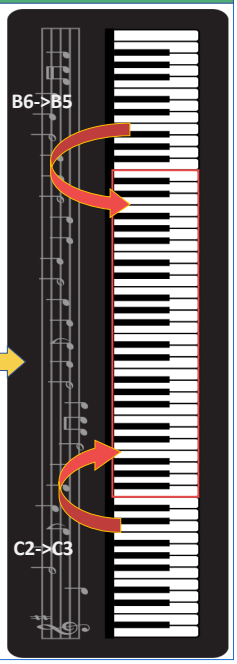
Yingfeng Fu[1], Yusuke Tanimura[2,1], Hidemoto Nakada[2,1]
The University of Tsukuba[1], The National Institute of Advanced Industrial Science and Technology (AIST)[2]

Abstract

- We try to learn the latent representation of different genres music. And improve the work based on[1].
- We modified the structure, keeping the same structure for generation network encoder, discriminator network and classifier. And we have a novel data compression.
- We successfully decrease the training time a lot, but the generated music quality still in evaluation progress.

Data preprocessing

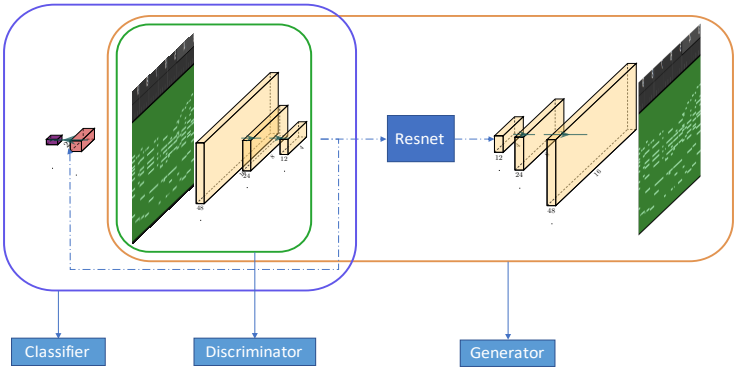




Pre-selection:
- Choose MIDI with drum tracks<=3, and total melody octaves<=5

Compression:
- Remove drum tracks
- Merge the remaining tracks into one :
- Fold in outlier pitches by an octave :

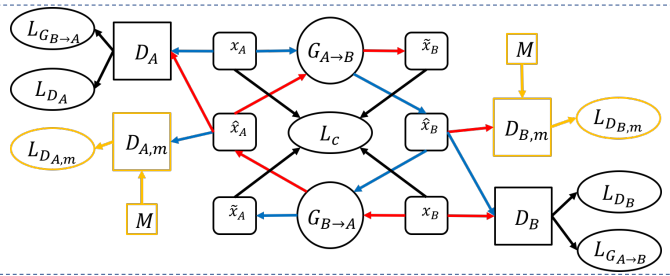
Model



Notations： X_A is the real A genre music data , $\hat{X}_B$ is the same data transferred to B genre , and $\hat{X}_A$ is the data goes back from the cycle. M is mixed multi-genre music data.

$$L_{G_{A\to B}} = \|D_B(\hat{X}_B)1-1\|_2 \qquad L_{G_{B\to A}} = \|D_A(\hat{X}_A)1-1\|_2 \qquad L_C = \|\tilde{X}_A - X_A\|_1 + \|\tilde{X}_B - X_B\|_1$$

$$L_G = L_{G_{A\to B}} + L_{G_{B\to A}} + \lambda L_C$$



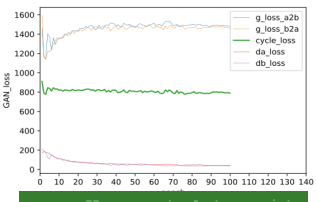
$$L_{D_A} = \frac{1}{2}(\|D_A(X_A)-1\|_2 + D_A(\hat{X}_A))_2 \qquad L_{D_B} = \frac{1}{2}(\|D_B(X_B)-1\|_2 + D_B(\hat{X}_B))_2$$

$$L_{D_{A,m}} = \frac{1}{2}(\|D_{A,m}(X_m)-1\|_2 + D_{A,m}(\hat{X}_A))_2 \qquad L_{D_{B,m}} = \frac{1}{2}(\|D_{B,m}(X_m)-1\|_2 + D_{B,m}(\hat{X}_B))_2$$

$$L_{D,all} = L_D + \lambda(L_{D_{A,m}} + L_{D_{B,m}})$$

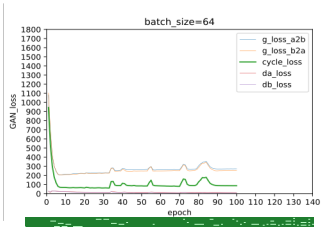
Experiment And Result

Experiment	Conv1,Deconv3	stride	Other CNN	stride	100 epochs
1[1]	7*7	1*1	3*3	2*2	10h
2	16*12	1*1	4*4	2*2	8.5h
3	16*12	1*1	3*3	2*2	8.5h
4	8*12	1*1	3*3	2*2	9h
5	16*12	4*1	3*3	2*2	2h

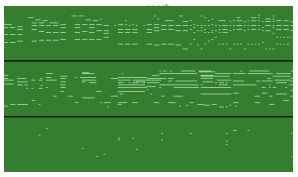


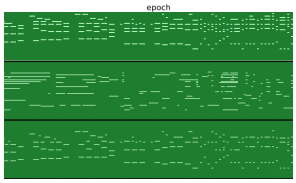
Left loss plot is using their model and our data representation

Right loss plot is using their model and our data representation



MIDI screen shot
Left their, right ours
Top: classic
Middle: to jazz
Bottom: cycle to classic





The table shows our data representation + model structure significantly reduced the training time. The output midi shows similar quality as their model. We believe the bigger non-square receptive field have a better vision in terms of the heterogeneity of data.

Conclusion And Discussion

- We improve the training speed and maintain the generation music performance. But both still not beautiful enough to be listened to.
- The generated music can't be recognized well by genre. And our classifier is in still being optimized.
- GAN training is hard, the good loss plot does not guarantee good output.
- The symmetry kept in our model+our data is better(less sparser output).

Reference

[1]Brunner, Gino et al. "Symbolic Music Genre Transfer with CycleGAN." *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)* (2018): 786-793.

Acknowledgement

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was supported by JSPS KAKENHI Grant Number 19K11994.