Bach in Black: Music Style Transfer

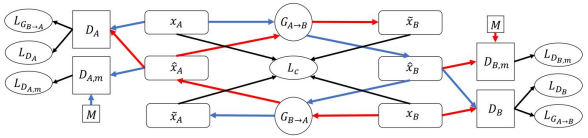
Cheng-You Lu, Hill Chang, Tiffany Zhang

Introduction

The goal of our project is music style transfer, creating a model that can take in as input a song from one genre of music and output the same song in a different genre. There is, however, a key difference between this problem and language translation: we do not have paired data. The dataset does not have examples of the same song in different genres. As such, our model has to learn by itself the features that identify a genre.

Background

One common approach to this task is to use Generative Adversarial Networks (GANs). Specifically, Zhu et al. [1] developed the concept of Cycle-Consistent Adversarial Networks for image-to-image style transfer, which uses two GANs that feed into each other. Brunner et al. [2] introduced CycleGAN to the music style transfer model, using ResNet [3] as their backbone with CNN-based GANS and achieved noteworthy results. Their model architecture can be seen below.



We think that using the CycleGAN architecture makes sense for music style transfer, but we feel that convolutional neural networks (CNNs) are not necessarily the best option when it comes to representing sequential data like music. To overcome this issue, we aim to substitute the backbone with a long short-term memory (LSTM) [3] RNN as the base for our model.

Data

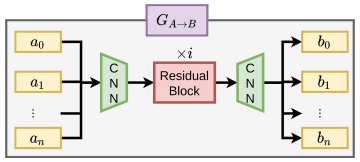
We used the already preprocessed 2,714 classical songs and 1,069 jazz songs from the dataset in [2]. The data was collected from a MIDI music database and then processed as follows:

- Cut off octaves that are rarely used
- Squeeze instruments into one piano track
- Broken into 4-measure sections

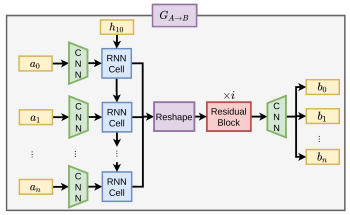
To generate audio that humans can listen to, we reversed this procedure.

Methodology

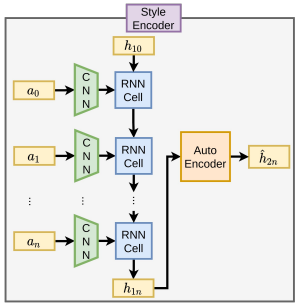
We first implemented the CNN-based CycleGAN model from [2] as a basis for comparison. We then designed our own LSTM-based generator to be used. In addition, we decided to add a style encoder. During training, this takes as input a random song from the target genre and encodes its style, to be used as the initial hidden state. We designed two versions of the LSTM-based generator. The first version V1 had multiple RNN layers, but ran into mode collapse issues. The second version V2 , whose architecture is shown below, uses larger time steps and fewer layers, and was able to achieve successful results without collapse.



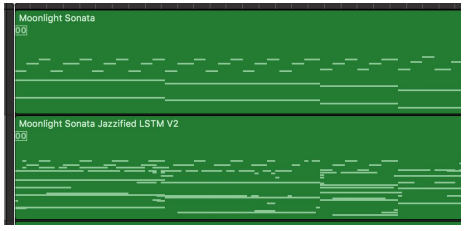
CNN-based generator



RNN-based generator



Style encoder



Visual comparison of MIDI songs. Top: Original song. Bottom: Model output

References

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in ICCV, 2017.
[2] Gino Brunner, Yuyi Wang, Roger Wattenhofer, and Sumu Zhao, "Symbolic music genre transfer with cyclegan," in ICTAI, 2018.
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in CVPR, 2016.
[4] Sepp Hochreiter and J ̈urgen Schmidhuber, "Long short-term memory," in Neural Computation, 1997.

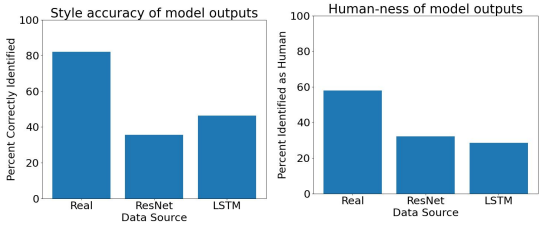
Results

We followed Brunner's method of calculating scores to measure the performance of our models between A:Jazz and B: Classical. First, we trained a classifier to classify the genre. We consider the style transfer a success if the source style is more likely before the transfer, and less likely after the transfer. Score was calculated using the following formula, and the results are in the table below.

$$S^D_{A\to B} = \frac{P(A|x_A) + P(A|\tilde{x}_A) - 2P(A|\hat{x}_B)}{2} \qquad S^D_{tot} = \frac{S^D_{A\to B} + S^D_{B\to A}}{2}$$

Pipeline	CNN	LSTM V1	LSTM V2	Base (Zhao)
A	82.86	82.86	82.86	88.09
$A \to B$	3.83	5.35	23.13	20.62
$A \to B \to A$	81.88	64.20	75.54	88.18
B	93.84	93.84	93.84	92.53
$B \to A$	9.82	15.98	14.91	20.71
$B \to A \to B$	94.32	91.70	94.11	92.53
S^D_{tot}	81.32	72.48	67.57	69.7

The performance measured on classifier in terms of A:Jazz and B:Classical



Survey Results

Discussion

Lingering Problems

Our best model LSTM V2 output was still identified as the original genre the majority of the time (53.6%) by humans, showing we did not transfer enough. The output also sounds very mechanical due to the aggressive simplifications in the MIDI representation and preprocessing.

Future Work

- Explore transformer more in depth
- Try variational autoencoder instead of autoencoder to encode style
- Use more complex representations of music and try different genres