GANtastic: Music Style Transfer with CycleGAN

Ganesh Arkanath
napoom@iu.edu

Meet Palod
mpalod@iu.edu

Aditya Ramachandra
ar83@iu.edu

Balajee Devesha Srinivasan
basrini@iu.edu

Abstract— **Several techniques for art style transfer have emerged with the most recent advancements and improvements in the fields of generative AI. One of the more innovative approaches involves leveraging Generative Adversarial Networks (GANs). When applying the GANs to the domain of artistic style transfer, there is a need for paired data, each representing the same context in different artistic domains. However, in the realm of music style transfer, paired data for the same example across different domains is rare or usually non-existent, for e.g. a song written in the classical genre seldom has a jazz equivalent. To overcome this challenge, the CycleGAN architecture employs 2 generator-discriminator pairs to maintain the original structure of the piece while incorporating changes specific to the target domain. In our implementation, we use a CycleGAN with 2 Generators and 2 Discriminators to achieve cycle consistency. Our dataset consists of unpaired music genre-specific piano arrangements from "The Lakh MIDI piano dataset". Our work aims to successfully convert the structure of the input MIDI from one domain (genre: Classical) to another domain (genre: Jazz) without losing its identity completely. For the experiments we conduct, we can demonstrate the applicability and potential of our methodology in generating non-trivial changes to the structure of the music data.**

Keywords—Generative AI, Deep Learning, Style Transfer, Music Transformation, CycleGAN, MIDI

I. Introduction

Deep learning has seen an unprecedented convergence of art and technology, pushing limits and giving rise to novel applications that engage our senses. By transferring stylistic elements between pieces with ease, this merging of disciplines has unlocked the ability to transform musical landscapes and inspire a new generation of artistic expression.

While the concept of music style transfer promotes innovation, previous methodologies frequently rely largely on paired datasets. These strictly curated sets, where each input sample has a corresponding target style, are limited and restrict the exploration of artistic possibilities.

Our project addresses this challenge by leveraging CycleGAN, a type of GAN that has shown remarkable success in style transfer tasks with unpaired data. CycleGANs introduce a cycle consistency loss to enforce the intuition that if we translate from one domain to another and back again, we should arrive at where we started. This makes them particularly suited for tasks like music style transfer, where paired data is rare or usually non-existent.

This project delves into the innovative fusion of CycleGAN with music style transfer, transferring piano tunes from Classical to Jazz. By embracing unpaired datasets, we aim to transcend the limitations of traditional methods. Our exploration will not only broaden our understanding of artistic expression but also unravel the latent capabilities of neural networks in reshaping auditory aesthetics.

This is a particularly interesting problem as it holds significant promise for other applications in the audio domain. This work can be extended to music restoration, speech enhancement, audio source separation, voice conversion, and more.

These are just a few examples of the vast potential of CycleGAN in the audio domain. By harnessing its versatility, we can unlock new avenues for creative expression, enhance the quality of audio experiences, and unlock new possibilities for audio processing and manipulation.

II. Literature Survey

Gatys et al. [1] developed the concept of neural style transfer, which shows how pre-trained convolutional neural networks (CNNs) can merge the style and content of images. Conversely, methods such as CycleGAN [2] eliminate the need for explicit feature extraction because they use two generators to transmit style between different domains in an elegant manner. For example, CycleGAN does not need to actively learn content and style features to convert horse images into zebra-style images and vice versa. We are extending this concept to MIDI music, to demonstrate CycleGAN's potential for style transfer, specifically in the field of genre, within symbolic music.

CycleGAN's underlying theory has changed dramatically over time, leading to the development of notable instances like MelGAN-VC [3], StarGAN [4], CoGAN [5], and DualGAN [6]. As we demonstrate the usefulness of CycleGAN for style transmission, especially in symbolic music, we look forward to exploring more into these works' developments and advancements in the future. Previous work on music style transfer includes the work of Malik et al. [7], in which they presented a model that uses velocities in "flat" MIDI files to simulate human-like music performance. However, the nuances of different musical genres and styles are not captured by this model, which is limited to altering note velocities. In another work, Brunner et al. [8] created MIDI-VAE, a multitask Variational Autoencoder that can change a piece of music's style, like from jazz to classical. But, our method is not limited by the number of notes played simultaneously, unlike MIDI-VAE, which produces richer and more dynamic music output for a realistic style transfer. Furthermore, Mor et al. [9] developed a WaveNet [10] autoencoder-based model for translating raw music between instruments, genres, and styles, while Van den Oord et al. [11] presented a VAE model with a discrete latent space for speaker voice transfer.

Our project's main goal is to successfully transmit symbolic music styles between many genres, with listeners' enjoyment of the music serving as a gauge of success. Even if it has nothing to do with style or domain transfer, we investigate studies in this field to obtain a clear picture of music generation. Music modeling is achieved using a variety of techniques, such as CNNs, LSTMs, and RNNs ([12], [13]). Long-term dependencies in polyphonic music have been captured by generative models such as GANs and VAE, with MusicVAE [14] being a notable hierarchical VAE model. The usefulness of CNN-based GANs for music creation has been demonstrated by recent work by Mogren [15], Yang et al. [16], and Dong et al. [17], despite the training issues associated with GANs. Inspired by Yu et al.'s [18] successful application of RNN-based GANs to music and their incorporation of reinforcement learning techniques, we use CNN-based GANs in our work to model music and enable domain transfer.

III. Dataset and preprocessing

In our experimentation, we are utilizing The Lakh MIDI Dataset as the primary source for our musical data. However, the raw data is not suitable for our usage as it contains a lot of additional data that can interfere with the learning process. To tailor it to our use case, we employ a preprocessing pipeline tailored to our specific requirements. Firstly, we isolate the piano channel, removing all other instrumental channels. This step ensures that the piano arrangements are the only input, providing a clear and consistent dataset. Subsequently, we split the complete MIDI inputs into manageable 30-second chunks. This achieves efficient data management and uniformity across all samples which is required for the CNN.

Following segmentation, we convert these MIDI segments into interpretable piano rolls using the MIDI utils library, translating the MIDI data into a visually intuitive and analytically useful representation. The piano roll provides a two-dimensional visualization, where the x-axis represents time and the y-axis denotes pitch, offering a clear overview of the musical structure. Finally, these piano rolls are stored as numpy arrays with the .npy extension. This format ensures compatibility with various Python-based data processing and machine learning libraries, allowing seamless loading during training and inference.

Our implementation requires music from diverse genres (at least two) to effectively achieve style transfer. For this purpose, we use the collection of Jazz, Classical, and Pop musical styles from the lakh piano dataset. To maintain data quality and consistency, we did additional data cleaning involving discarding MIDI files that deviate from certain key criteria, like not starting on the first beat or featuring non-standard time signatures. This ensures the data is focused on standardized musical structures [Symbolic Music Genre Transfer with CycleGAN].

Following this initial filtering, our dataset comprises:

- 12,341 *Jazz samples*: Capturing the unique improvisational and syncopated nature of Jazz.
- 16,545 *Classical samples*: Encompassing the rich melodic and harmonic structures of Classical music.
- 20,780 *Pop samples*: Representing the diverse and dynamic sounds of Pop music.

There are several other advantages to using MIDI format for music style transfer compared to WAV or actual music files:-

Noise Elimination: MIDI data encodes musical information directly, eliminating the need to process and analyze audio recordings. This bypasses the challenges of background noise, ensuring a clean and consistent representation of the musical content.

Phase Reconstruction: Unlike WAV files, which require reconstructing the phase information during spectrogram analysis, MIDI data inherently contains timing information. This eliminates the need for complex and error-prone phase reconstruction algorithms, leading to a simpler and more robust analysis and manipulation of musical structure.

Flexibility and Control: MIDI data allows for precise control over individual notes, including pitch, velocity, and timing.

Compactness: MIDI files are significantly smaller compared to WAV files representing the same musical information. This translates to reduced storage requirements, faster processing times, and efficient data handling, which are crucial for large-scale music analysis and generation tasks.

Interoperability: MIDI is a widely supported and standardized format, facilitating seamless integration with various music software and tools. This interoperability enables researchers and musicians to leverage existing resources and libraries, promoting collaboration and innovation in the field of music style transfer.

IV. Model Architecture

A Cycle-GAN model is implemented, where we use 2 discriminators(D1 and D2) and 2 generators(G1 and G2) to create a cyclic architecture. Data from domain A data is fed to G1 to get a prediction in domain B. The generator is trained using four losses.

1. Adversarial loss - The output is G1 is tested against D1 to check whether the output falls in domain B.
2. Identity loss - A sample from domain B is fed to G1 and it is expected that the output is still in domain B. This helps to prevent the generator from introducing unwanted artifacts or distortions into the translated image
3. Forward cycle loss - This loss measures the difference between the original input (X) and the image reconstructed back from the translated image (G2(G1(X))). The forward cycle loss ensures that the translated image (G1(X)) retains the essential information from the original image and can be translated back without significant content loss.
4. Backward cycle loss - This loss measures the difference between the original image in the target domain (Y) and the image reconstructed back from the translated image (G2(G1(Y))). The backward cycle loss ensures that the translated image (F(Y)) retains the essential features of the target domain image and can be translated back without losing necessary details.

For adversarial loss, we use mean squared error and for other three losses, we use mean absolute error, so that we can perform a pixel wise comparison.

We assign weights to each loss, so as to ensure that the generator prioritizes certain aspects of the translation over others. It also helps prevent overfitting from any single loss, a common problem faced with CycleGans.

Whereas, discriminators are trained independently, by feeding them the original samples from domain A and the images generated from generator.

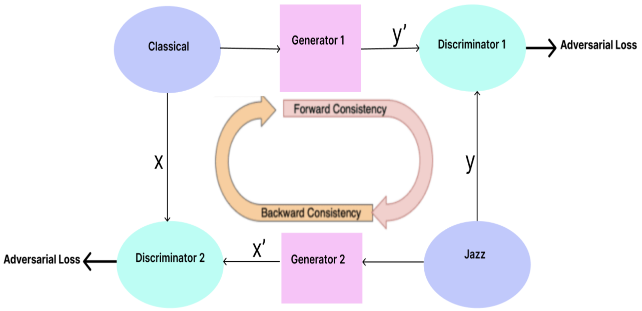
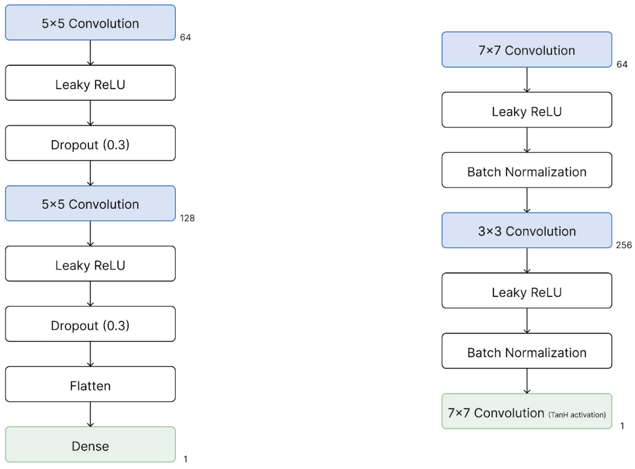


Fig.1 Cycle-GAN architecture

Discriminator Architecture - The discriminator takes input in the form of a 3D tensor (akin to an image of the midi file's corresponding piano roll representation, 3D excludes the batch dimension). The model comprises two convolutional blocks and a dense layer. Each convolutional block consists of a convolutional layer with 5x5 kernels and stride size of 2x2, followed by a Leaky ReLU activation with alpha=0.3, followed by a Dropout layer with rate=0.3. The output of the last convolutional block is flattened and fed to a dense layer with a single unit. The output of the discriminator is compared with a vector of 1s (for a sample of the original genre) or a vector of 0s (for a sample of the other genre) and loss is computed on this.



(a) Discriminator (b) Generator

Fig.2 Generator and Discriminator architectures

Generator Architecture - The generator takes input in the form of a 3D tensor (akin to an image of the midi file's corresponding piano roll representation, 3D excludes the batch dimension). The model comprises 3 blocks. The first block consists of a convolutional layer with 7x7 kernels, followed by LeakyReLU activation with alpha=0.3, followed by Batch Normalization. The second block consists of a convolutional layer with 3x3 kernels, followed by LeakyReLU activation and Batch Normalization. The last block is made up of a convolutional layer with a single 7x7 filter, followed by tanh activation.

V. Results and Conclusion

In our approach of symbolic music style transfer, we proposed a model that can successfully perform transfer style across different genres without relying upon paired data. The cycleGAN architecture yields impressive results, generating highly realistic samples.

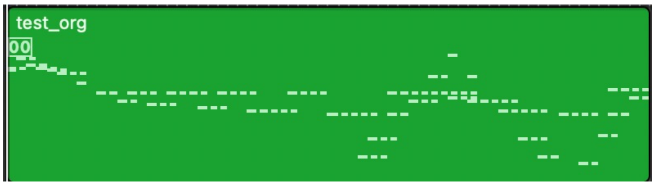


Fig.3.1 MIDI output of original file

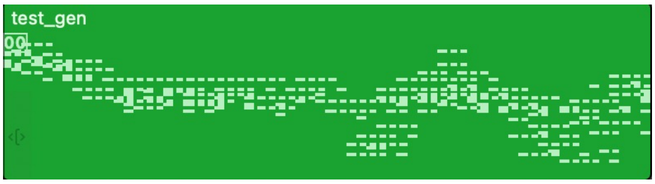


Fig.3.2 MIDI output of generated file

From Fig.3, there is a notable difference between the generated output and the original sample in terms of scale and chords of the musical notes, while the overall progression is similar, which are also distinguishable to the human ear. The generated sample of our model is also very pleasing to the listeners and generally sounds very harmonic since they do not include noises, as opposed to outputs generated by models leveraging spectrograms. Furthermore, these samples preserve the original music style while changing the music genres as we constantly monitor the cycle-loss which can be seen in Fig. 3. We also trained the model for 500 epochs and the best conversion is observed around 200 epochs, while preserving the style of both the genres. Overtraining the model leads to the loss of original tune in the output and is more close to jazz.

VI. Future Work

The project has the potential to significantly increase the effectiveness of symbolic music style transfer by using some cutting-edge architectures. With the proven dominance of ResNext blocks over ResNet, leveraging them in our model could significantly improve the models performance. Also, adopting multi-class classifiers for discriminator, as opposed to binary classification, offers more flexible and robust comprehension of music styles. Furthermore, exploring some complex GAN architectures, such as ReCycle GAN and BiCycle GAN, has the potential to effectively transfer style across different domains. These developments could not only improve the performance of our existing model but also expand the boundaries of music generation.

References

[1] Gatys, L.A., Ecker, A.S. and Bethge, M., 2016. Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).

[2] Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

[3] Pasini, M., 2019. MelGAN-VC: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms. arXiv preprint arXiv:1910.03713.

[4] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S. and Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797).

[5] Liu, M.Y. and Tuzel, O., 2016. Coupled generative adversarial networks. Advances in neural information processing systems, 29.

[6] Yi, Z., Zhang, H., Tan, P. and Gong, M., 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE international conference on computer vision (pp. 2849-2857).

[7] Malik, I. and Ek, C.H., 2017. Neural translation of musical style. arXiv preprint arXiv:1708.03535.

[8] Brunner, G., Konrad, A., Wang, Y. and Wattenhofer, R., 2018. MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer. arXiv preprint arXiv:1809.07600.

[9] Mor, N., Wolf, L., Polyak, A. and Taigman, Y., 2018. A universal music translation network. arXiv preprint arXiv:1805.07848.

[10] Oord, A.V.D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K., 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

[11] Van Den Oord, A. and Vinyals, O., 2017. Neural discrete representation learning. Advances in neural information processing systems, 30.

[12] Todd, P.M., 1989. A connectionist approach to algorithmic composition. Computer Music Journal, 13(4), pp.27-43.

[13] Mozer, M.C., 1994. Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. Connection Science, 6(2-3), pp.247-280.

[14] Roberts, A., Engel, J. and Eck, D., 2017, December. Hierarchical variational autoencoders for music. In NIPS Workshop on Machine Learning for Creativity and Design (Vol. 3).

[15] Mogren, O., 2016. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. arXiv preprint arXiv:1611.09904.

[16] Yang, L.C., Chou, S.Y. and Yang, Y.H., 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. arXiv preprint arXiv:1703.10847.

[17] Dong, H.W., Hsiao, W.Y., Yang, L.C. and Yang, Y.H., 2018, April. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

[18] Yu, L., Zhang, W., Wang, J. and Yu, Y., 2017, February. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1)