Music Style Transfer with Vocals Based on CycleGAN

Hongliang Ye and Wanning Zhu
Department of Software Engineering, Jinling Institute of Technology, Nanjing 211100, China
Email: zhuwanning@jit.edu.cn

Abstract. In recent years, with the development of generative adversarial networks (GAN), the application of generative adversarial networks has gradually matured. An important application area for generating adversarial networks is called neural style transfer. In recent years, neural style transfer has played a major role in the field of image applications. However, it performed poorly in the music field. In addition, algorithms in the field of music style transfer have poor effect on the style transfer of music with vocals. Therefore, this paper extracts the CQT features and Mel spectrogram features of music, and then uses CycleGAN to transfer the styles of the CQT features and Mel spectrogram mapping pictures, and finally realizes the style transfer of music. On the classifier we trained, the average style transfer rate of music that meets our requirements reached 94.07%.

1. Introduction

1.1. Background
Over the past three years, neural style transfer has continued to grow into a thriving research area [1]. More and more activities in this research field are driven by scientific challenges and industrial needs. Style transfer has broad application prospects including social networking, assisting user creation and entertainment applications.

Music style transfer is another attempt in the field of style transfer algorithms [2]. Since music is a time-based segment and there are many music components, feature extraction is more complicated, and the connection between features is more complex and tight. At present, most researches apply image style transfer algorithms directly to music style transfer, and most of the music is pure music. However, the results of these algorithms in popular music with vocals are not satisfactory. At present, a large number of songs are covered in various versions of different styles, but the number of singers’ cover is far from meeting people’s needs for different styles of cover songs. Therefore, it is of great significance for the computer music field to study a model suitable for the transfer of popular music styles with vocals.

1.2. Related Work
The work done by Gaty et al. [3] used neural networks for image style transfer for the first time and produced surprising results. After this, some people applied Generative Adversarial Network (GAN) in the field of image style transfer. Subsequently, the field of style transfer gradually developed in the field of computer vision. This also promoted the vigorous development of GAN. In recent years, unsupervised learning GANs such as CycleGAN [4], DualGAN [5], DiscoGAN [6] have been
gradually proposed. Usually, it is difficult to find paired data such as pictures and music. These several unsupervised algorithms have solved these problems.

At present, the work done by researcher on the transfer of musical styles is mainly in the transfer of pure musical tone styles of common musical instruments. Gino et al. [7] converted midi format audio into a piano rolling matrix, then input the matrix into CycleGAN for training, and finally generated transferred midi audio. Timbretron proposed by Huang et al. [8] extracts the CQT features of audio, and then converts its timbre through CycleGAN. Then convert the converted CQT features into the original audio. Noam et al. [9] proposed a universal music translation network. This network implements music timbre transfer by training a WaveNet music encoder and multiple WaveNet decoders. The network can realize the transfer from one timbre domain to multiple timbre domain. The above-mentioned algorithms have shown good results in their respective research directions, but in the literature we have read, we have hardly seen any research on the style transfer of audio with vocals and background music. Directly forcing their algorithm onto this problem did not produce a convincing effect. Therefore, this paper extracts the CQT features of the audio and the Mel spectrogram, and then uses CycleGAN to transform the features into styles, and finally converts the features into high-quality music through the WaveNet decoder, and finally realizes the style transfer of music with vocals.

2. Previous Work

2.1. CycleGAN

Generative adversarial network is a kind of implicit generative model proposed by Goodfellow et al. [10]. The model produces a fairly good output through the mutual game learning of the generative model and the discriminant model. The generative model attempts to generate fake samples to fool the discriminant model. The discriminant model attempts to distinguish real data from fake samples. Suppose $G$ is a generator, $D$ is a discriminator, $P_{data}(x)$ is the distribution of real samples and $x$ is sampled from the distribution, and $P_z(z)$ is the distribution of the latent code $d$ of $x$. Then the objective is:

$$G, D = \arg \min_G \max_D E_{x \sim P_{data}} [\log D(x)] + E_{z \sim P_z} [\log (1 - D(G(z)))]$$

(1)

CycleGAN is an unsupervised generative adversarial network. CycleGAN can learn the mapping between two domains without any paired data. CycleGAN contains two generators and two discriminators. The two generators need to learn the mapping of this domain to the corresponding domain respectively. The two discriminators need to learn whether the data generated by the corresponding domain generator is local domain data by learning the real data of the respective domains. In addition to two adversarial losses, The loss function in CycleGAN also needs to add a cycle consistency loss to preserve its input structure. The loss of cycle consistency is defined as follows:

$$L_{cycle}(G, F) = E_{x \sim P_{data}} [\| F(G(x)) - x \|_1] + E_{y \sim P_{data}} [\| F(G(y)) - y \|_1]$$

(2)

2.2. Constant Q Transform

In mathematics and signal processing, the constant Q transform transforms the data sequence into the frequency domain. This transformation reflects the human auditory system, so that the spectral resolution is better at lower frequencies and the temporal resolution is improved at higher frequencies. In addition, in this transformation, the harmony of the notes forms a pattern characteristic of the timbre of the musical instrument. Assuming that the relative intensity of each harmonic is the same, as the fundamental frequency changes, the relative position of these harmonics will remain constant. This can make the model identification easier.
2.3. Mel-Spectrogram
Since the spectrogram obtained after Fourier transform is large, in order to obtain sound features of appropriate size, it is usually passed through the Mel scale filter bank to become the Mel spectrogram. The human ear is not linearly sensitive to the frequency of sound, but is sensitive to low frequencies and insensitive to high frequencies. When the spectrogram is converted into the Mel spectrogram, the human ear’s perception of frequency becomes linear. This makes it easier to extract and process feature processing.

2.4. WaveNet
WaveNet, proposed by van den Oord et al. [11], is an autoregressive generation model for generating high-quality original audio waveforms. The model consists of an extended causal convolutional layer with residual connections and skip connections. WaveNet adopts the idea of causal convolution and dilated convolution. Causal convolution ensures that the network will not obtain advanced information when predicting, and dilated convolution can increase the receptive field of the network, and the two make the generated audio more nature.

3. Model Architecture
In this chapter, we introduced the basic architecture of our model. Our model is based on CycleGAN and WaveNet decoders. The model processing flow is as follows and:

- First extract the Mel spectrogram features and CQT features of the audio.
- Then merge the extracted Mel spectrogram features and CQT features into two layers and input them into the CycleGAN model. Then CycleGAN generates the converted Mel spectrogram features and CQT features.
- Input these two layers into the WaveNet decoder trained beforehand to generate audio. Figure 1 shows our model structure:

![Figure 1. Model architecture.](image)

3.1. CycleGAN

3.1.1. Losses Used. Because the goal of the algorithm is to transfer music from one domain to another, the generator does not actually take noise as input, but obtains real samples from the source domain. In this article, we only deal with translation between two domains at a time, so we call them X and Y, which correspond to music from two different genres. Since the transmission should be symmetrical, that is, we want to transfer samples from dominax to domainy and vice versa. The basic loss function of CycleGAN( X → Y ) is shown in equation (3):

\[
L_{GAN}(G, F, X, Y) = E_{x \sim P_{data}(x)}[\log D(x)] + E_{y \sim P_{data}(y)}[\log (1 - D(G(y)))]
\]

(3)

In addition, we also added the identity loss proposed by Zhu et al., because the experiment shows that when this loss is not added, the generated spectrogram will lose its color component, which is manifested as making the final generated audio indirectly produce a larger Noise, so adding this loss will help the model generate a higher quality spectrogram.
\[ L_{\text{identity}}(G, F) = E_{x \sim p_{\text{data}}(x)}[\| G(y) - y \|_1] + E_{y \sim p_{\text{data}}(y)}[\| F(x) - x \|_1] \]  

(4)

Add cycle consistency loss and identity loss expressed by equation (2), where \( \lambda_1 \) and \( \lambda_2 \) respectively represent the weight of cycle consistency loss and identity loss, so the total loss function is:

\[ L(G, F, D_X, D_Y) = L_{\text{GAN}}(G, D_Y, X, Y) + L_{\text{GAN}}(F, D_X, Y, X) + \lambda_1 L_{\text{cycl}}(G, F) + \lambda_2 L_{\text{identity}}(G, F) \]  

(5)

The weights \( \lambda_1 \) and \( \lambda_2 \) of cycle consistency loss and identity loss will significantly affect the generated audio. For \( \lambda_1 \), when \( n_1 \) is too large, CycleGAN will choose a simple, low-latitude transformation, and when \( \lambda_1 \) is too small, CycleGAN will seek a transformation with a higher complexity, and the resulting transformation styles are diverse but not easy to control. As far as \( \lambda_2 \) is concerned, when \( \lambda_2 \) is too large, CycleGAN will place strong constraints on the original style, which will make the generated audio lose its expressiveness. When \( \lambda_2 \) is too small, CycleGAN will lose the color components of the spectrogram. Since the randomly changing value will make the algorithm unable to converge normally, this paper sets \( \lambda_1 \) as a fixed value in the experiment. For \( \lambda_2 \), research shows that as the number of algorithm iterations increases, the value of should be smaller [8]. Usually the attenuation of \( \lambda_2 \) is linear attenuation, but this paper found through experiments that the curve attenuation is more in line with the music style transfer algorithm. Therefore, this paper attempts to propose a non-linear attenuation function for \( \lambda_2 \). Compared with the linear attenuation function, the non-linear attenuation function proposed in this paper makes the model exhibit better robustness. Assuming that the algorithm needs to iterate \( t \) steps in total, then at the \( n \) The \( \lambda_2 \) value is:

\[ \lambda_2 = \frac{t + \sqrt{t}}{n + \sqrt{t}} \]  

(6)

3.1.2. Checkerboard Artifacts. Since the CycleGAN of Zhu et al. [4] uses a deconvolution operation, this will cause the generated spectrum to have serious checkerboard artifacts, and the impact on the audio will be shown as serious indirect noise (as shown in the red box in figure 2). To this end, we refer to the work of Odena et al. [12], using nearest neighbor interpolation and regular convolution instead of deconvolution. This method works well in image super-resolution, but because it does not use deconvolution operation, it is not easy to have the result of checkerboard artifacts (figure 3).

![Figure 2. Audio generated using deconvolution.](image1)

![Figure 3. Audio generated using nearest neighbor interpolation.](image2)

3.2. WaveNet Decoder

3.2.1. Input Data Preprocessing. Because WaveNet loss function is tanh loss function. The waveform represented by the tanh function is between [-1,1], and the values of the natural logarithm of the CQT spectrogram generated by audio and the Mel spectrogram conform to the normal distribution between
Therefore, we standardize the input data globally so that the distribution of the input data conforms to the N(0,1) distribution. Because it is difficult to predict the phase directly from the spectrogram in the time-frequency representation [8], so we drop the phase information of the Mel spectrogram and the CQT spectrogram, and then deform and zero-fill the two-layer phase information to facilitate merging the two layers.

3.2.2. Network Architecture. In this article, the WaveNet decoder has an expansion rate of $2^k$ (k indicates which layer the network is in). For all dilated convolution and causal convolution layers, we use a convolution kernel of size 3. For all residual blocks, the length of the skip connections and the residual connections are 256. In addition, each residual layer contains a ReLu function.

4. Experiment
Evaluating the performance of a music generation system is difficult, because the quality of music is a highly subjective measure. Evaluating styles and domain transfer is a bit simpler, because a classifier can be trained based on the existing training set. Then determine the transfer rate of style transfer based on the results of the classifier. Based on the above two points, this article has adopted the following scheme:
- According to people’s subjective evaluation of whether the music quality is qualified, such as whether there is a large noise.
- Train a classifier based on the existing music data to determine the success rate of the music domain.

4.1. Datasets
The dataset selected by our experiment is FMA [13], which is an open and easy-to-access dataset, suitable for research analysis such as music classification. For the choice of the size of the dataset, we chose the medium version. This version contains a total of 25,000 music clips, each music clip has a length of 30s and a sampling rate of 22050. For the music genres in the dataset, we selected the following 6 categories for experiments. These include Pop, Blues, Folk, Jazz, Country, Classical. Due to network limitations, we divided a 30s segment into six approximately 5s segments and input them into CycleGAN for style transfer. Extract CQT features and Mel spectrogram features through librosa [14]. Then CQT and Mel spectrogram are merged into two layers, and the parts that are not aligned are filled with 0.

4.2. Classifier Training
Referring to the work of Wu et al. [15] in music genre classification, we trained the DW-KNN music genre classifier on the FMA dataset. The included music domains are as follows, and the accuracy of various predicted music genres is shown in the following table 1.

| Music Type | Pop  | Blues | Folk | Jazz  | Country | Classical |
|------------|------|-------|------|-------|---------|-----------|
| Accuracy   | 69.82| 74.02 | 62.16| 69.19 | 50.88   | 56.63     |

4.3. Transferred Music Quality Assessment
Since the generated audio may not be guaranteed in quality, we randomly check multiple music clips in the transfer results in each domain. In addition, We arrange testers to make quality judgments on the results of spot checks. For specific indicators and results, see Section 4.4.

4.4. Music Style Transfer Assessment
Because there are many types of music genres, the boundaries are blurred, and the music has a time characteristic, a song may contain multiple segments of different styles. In order to better judge the
effect of this style transfer algorithm, we abandoned the subjectivity. The transferred music style is marked according to the classification result of the classifier. Among the multiple sets of data in the experiment, we selected the following 6 sets and their inverse processes for evaluation. Evaluation indicators include Audio Quality Rate (AQR), Transfer Rate (TR).

First, randomly select 30 music clips from each sample of the converted music domain, where AQR is determined by the average score of 5 people on the audio. First, the scorer evaluate the transferred music samples of each domain, and then rate audio. The score range is 0~5. The closer the AQT score is to 5, the higher the quality of the music is. TR indicates that the classifier is used to judge the proportion of the music that has undergone style transfer and successfully transferred to the new music domain.

From tables 2 and 3, it can be seen that CycleGAN’s style transfer rate (TR) has significantly improved when the audio quality is not much different. The forward style transfer rate (TR) has increased by an average of 2.52%, and the backward transfer rate (TR) increased by an average of 1.85%, so λ adopts the non-linear attenuation scheme proposed in this paper to show better style transfer ability than linear attenuation. To some extent, the average style transfer rate of our algorithm has reached 94.07%, showing a good representation learning ability, and achieving the style transfer of the above 6 music genres. In addition, because FMA contains a lot of music with vocals, the average audio quality pass rate (AQR) of our algorithm after transfer has also reached 82.86%, so in the processing of music with vocals, our algorithm also shows good results and has certain robustness.

| Origin domain ->Object domain | Forward transfer | Backward transfer |
|-----------------------------|-----------------|------------------|
| AQR(%) | TR(%) | AQR(%) | TR(%) |
| Pop->Classical | 86.23±5.65 | 93.66 | 80.17±6.64 | 92.34 |
| Pop->Blues | 87.16±4.82 | 96.74 | 75.91±3.31 | 94.75 |
| Blue->Country | 82.38±5.23 | 93.79 | 79.44±8.07 | 91.53 |
| Folk->Jazz | 88.69±5.50 | 89.45 | 78.08±8.20 | 88.03 |
| Jazz->Classical | 79.79±6.95 | 91.71 | 75.48±8.39 | 89.14 |
| Pop->Folk | 84.23±4.68 | 94.42 | 80.77±6.12 | 91.41 |

| Table 3. AQT (mean±SD), TR scores for music style transfer task with \( \lambda_i \) in equation (6). |
|-----------------------------|-----------------|------------------|
| Origin domain ->Object domain | Forward transfer | Backward transfer |
| AQR(%) | TR(%) | AQR(%) | TR(%) |
| Pop->Classical | 85.72±6.78 | 96.42 | 82.27±5.34 | 95.12 |
| Pop->Blues | 89.42±5.32 | 97.80 | 79.46±4.44 | 93.95 |
| Blue->Country | 86.91±7.43 | 95.59 | 79.71±6.87 | 94.32 |
| Folk->Jazz | 82.48±8.22 | 92.55 | 80.22±7.20 | 89.52 |
| Jazz->Classical | 81.79±4.35 | 95.30 | 78.42±5.25 | 91.74 |
| Pop->Folk | 86.32±6.62 | 94.73 | 81.64±6.86 | 91.82 |

5. Conclusion
In this paper, by extracting the CQT features and Mel spectrogram features of music, and then using CycleGAN to transfer the styles of the CQT features and Mel spectrogram feature maps, and finally generating the music waveform through the WaveNet decoder, the style of music with human voice is finally transferred with our model. On the classifier we trained, the average style transfer rate of music that meets our requirements reached 94.07%. This paper extracts the above two features, which not only retains the characteristics of music, but also retains the characteristics of more vocals. At the same time, the overlapping parts of these two features jointly constrain the waveform. This make WaveNet predict the waveform more accurately.
Acknowledgments
This article is funded by the Jinling Institute of Science and Technology, a high-level talent research startup fund (jit-b-201624), Jiangsu Educational Technology Research Institute, Modern Educational Technology Research Project (2019-R-79396) and the Jinling Institute of Technology's Innovation and Entrepreneurship Competition, Research on Popular Music Style transfer Based on Deep Learning (202013573045Y).

References
[1] Jing Y, Yang Y, Feng Z, et al. 2019 Neural style transfer: A review IEEE Transactions on Visualization and Computer Graphics.
[2] Dai S, Zhang Z, Xia G G. 2018 Music style transfer: A position paper arXiv preprint arXiv:1805.06841.
[3] Gatys L A, Ecker A S, Bethge M. 2015 A neural algorithm of artistic style arXiv preprint arXiv:1508.06576.
[4] Zhu J Y, Park T, Isola P, et al. 2017 Unpaired image-to-image translation using cycle-consistent adversarial networks Proceedings of the IEEE International Conference on Computer Vision pp 2223-2232.
[5] Yi Z, Zhang H, Tan P, et al. 2017 Dualgan: Unsupervised dual learning for image-to-image translation Proceedings of the IEEE International Conference on Computer Vision pp 2849-2857.
[6] Kim T, Cha M, Kim H, et al. 2017 Learning to discover cross-domain relations with generative adversarial networks arXiv preprint arXiv:1703.05192.
[7] Brunner G, Wang Y, Wattenhofer R, et al. 2018 Symbolic music genre transfer with cyclegan 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI) (IEEE) pp786-793.
[8] Huang S, Li Q, Anil C, et al. 2018 Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer arXiv preprint arXiv:1811.09620.
[9] Mor N, Wolf L, Polyak A, et al. 2018 A universal music translation network arXiv preprint arXiv:1805.07848.
[10] Gosodfellow I, Pouget-Abadie J, Mirza M, et al. 2014 Generative adversarial nets Advances in Neural Information Processing Systems pp 2672-2680.
[11] Oord A, Dieleman S, Zen H, et al. 2016 Wavenet: A generative model for raw audio arXiv preprint arXiv:1609.03499.
[12] Odena A, Dumoulin V and Olah C 2016 Deconvolution and checkerboard artifacts Distill doi: 10.23915/distill.00003.
[13] Defferrard M, Benzki K, Vanderheynst P, et al. 2016 FMA: A dataset for music analysis arXiv preprint arXiv:1612.01840.
[14] Librosa URL https://librosa.github.io.
[15] Wu M and Liu X 2020 A Double Weighted KNN Algorithm and Its Application in the Music Genre Classification 2019 6th International Conference on Dependable Systems and Their Applications (DSA) (IEEE) pp 335-340.