ChordGAN: Symbolic Music Style Transfer with Chroma Feature Extraction

Conan Lu[1] and Shlomo Dubnov[2]

[1] Redmond High School, Redmond, USA
[2] Department of Music, University of California San Diego
conanlu4@gmail.com

Abstract. We propose ChordGAN, a generative adversarial network that transfers the style elements of music genres. ChordGAN seeks to learn the rendering of harmonic structures into notes by embedding chroma feature extraction within the training process. In notated music, the chroma representation approximates chord notation as it only takes into account the pitch class of musical notes, representing multiple notes collectively as a density of pitches over a short time period. Chroma is used in this work to distinguish critical style features from content features and improve the consistency of transfer. ChordGAN uses conditional GAN architecture and appropriate loss functions, paralleling image-to-image translation algorithms. In the paper, pop, jazz, and classical datasets were used for training and transfer purposes. To evaluate the success of the transfer, two metrics were used: Tonnetz distance, to measure harmonic similarity, and a separate genre classifier, to measure the transfer style fidelity. Given its success under these metrics, ChordGAN can be utilized as a tool for musicians to study compositional techniques for different styles using same chords and automatically generate music from lead sheets.

Keywords: style transfer, chroma features, GAN, CNN

1 Introduction

The advent of generative adversarial networks (GANs) (Goodfellow et al., 2014) has introduced new accessible applications of machine learning, with many deeply intertwined with expanding and reimagining human expression. One such application is style transfer. Popular examples of style transfer are in the realm of image-to-image translation, in which an image in one domain is transformed to adopt the features and characteristics of another domain (Brunner, Wang, Wattenhofer, & Zhao, 2018). Pix2Pix expands on this concept by using conditional GANs (Isola, Zhu, Zhou, & Efros, 2018).

Style transfer is based on the principle of retaining some structural features from one domain, or type of data, while keeping the details, such as texture or other fine-grained information, from a second domain. The concept of style itself is not self-evident (Argamon, Burns, & Dubnov, 2010), and is often considered as

focusing on "how" things are done rather than "what" is done. In this spirit, we consider in this paper a specific structure aspect that is ubiquitous in music: the harmonic structure. By learning the relations between harmony, estimated via the chroma representation, and the musical surface,[1] we use the GAN framework to learn the relation between stylistic renderings of chords into notes.

The style captured with ChordGAN involved the rendering of chords within the chromagram representation into notes within the musical surface. Cambouropoulos et al. posit that chords do not appear strictly "below" the musical surface, but maintain a role within it (Cambouropoulos, 2016). The style transfer achieved with ChordGAN does not encompass the style involved in creating chords. Furthermore, ChordGAN does not capture long-term structure in music, instead focusing its relations to short-term voicing and melodic structure.

Early attempts at symbolic music style transfer include MIDI-VAE (Brunner, Konrad, Wang, & Wattenhofer, 2018), which utilized a variational autoencoder to learn musical style elements. MusicVAE (A. Roberts, Engel, Raffel, Hawthorne, & Eck, 2019) expanded on this by capturing long-term structure in polyphonic music. While setting an important precedent, many have cited the limitations, like the number of simultaneously played notes, as reasons for necessitating further development. MuseGAN (Dong, Hsiao, Yang, & Yang, 2017) and MidiNet (Yang, Chou, & Yang, 2017) both present the use of GANs for multi-track music generation. GAN architectures like CycleGAN have also been used for symbolic music style transfer; however, pre-trained convolutional neural networks were utilized to distinguish between content features and style features in symbolic music (Brunner, Wang, et al., 2018).

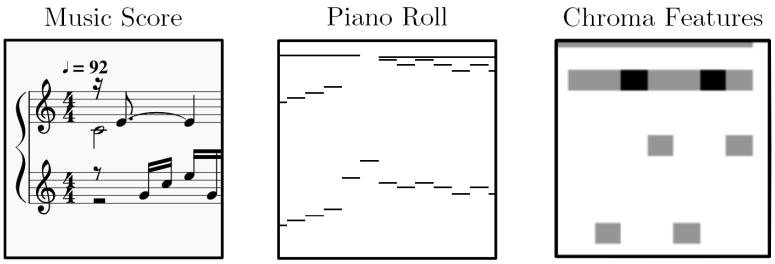

Fig. 1. Three formats (music score, piano roll, chroma features) for representing music. The chroma features only sample note density information from piano roll.

In our model, style transfer using GANs was achieved via pairing music, encapsulated in the piano roll format, with their corresponding chromagram

[1] By musical surface, we refer to the multitude of parameters representing notes and duration information contained in a MIDI file.

representation. The chromagram matrix is extracted from a piano roll by folding over octaves with dimensions {12, B}, where B signifies the number of bars (Wang & Dubnov, 2014). The chromagram representation format translates to a full piano roll in a manner analogous to the image-to-image translation of Pix2Pix. This medium of transfer ensures that the most important features are maintained, resulting in a realistic style transfer.

The first phase of the method trains a GAN to learn the style features of a particular genre or compositional style. The second phase actualizes the learned style features by facilitating a transfer. ChordGAN can receive any piece of tonal music as input and extracts the chromagram representation. The network then iterates through its translation, which will output that piece in the target domain.

The transfer results were evaluated using two metrics. The first was the Tonnetz distance, calculated between the two MIDI files before and after the transfer, to test the maintenance of content. Tonnetz is a harmonic representation of a piece of music in which lower distances correspond with more similar chordal structures. Using this metric as an indicator of success, it was found that the Tonnetz distance between all experimental transfers was 0.0. The distance between controls with random pairs was 0.0469. This indicates a successful maintenance of content. The second evaluator of the style transfer's performance was a separate genre classifier, as a convolutional neural network, to test the transformation of style. Based on music from real samples, the classifier had a predictive success rate of 80%. With the generated classical outputs as test data, the genre classifier had an average accuracy rate of 68.7% between genres.

From the success of the transfer in maintaining the content of musical pieces while also delivering convincing style changes, ChordGAN can be considered a successful style transfer system. One benefit of ChordGAN, compared to previous iterations, is the flexibility of the network to accommodate different genres of music. ChordGAN only controls the transfer of chroma features to a piece of music. Thus, any tonal music can be given as input to the network to generate a piece in the style of a particular genre.

ChordGAN as a method of style transfer presents several applications. Firstly, consistent music style is important for creating instrumental soundtracks. Because generative adversarial networks can apply transfers to large volumes of music relatively quickly, an effective transfer model would serve an important practical niche. Additionally, the ability to parse out distinguishing style elements of particular composers would be useful for music theory study. ChordGAN can be a useful tool for creative composition as a result. An additional application of our proposed method is rendering improvisations according to a chord progression or lead sheet. Current work centers on mapping chord notations to chroma features. This mapping must take into account a weighting scheme related to the voicing of a chord, adding more weight to harmonics or overtones belonging to the root and lower chordal notes and less to the embellishments or alterations in case of complex chords. We may also consider learning the chord to chromagram relation from lead sheet data in the future.

2 Model Architecture

Conditional GANs are comprised of the generator and discriminator networks to facilitate training. The generator G creates a realistic "fake" music score (represented via the piano roll) based on the chromagram representation z. This marks a departure from prior GAN architecture, in which z commonly denotes a random noise vectors. The discriminator receives either a generated sample $G(z)$ or real sample x as input, in conjunction with the initial chromagram representation z. The discriminator D attempts to distinguish between real and generated data. Fig. 2 demonstrates the architecture of the network.

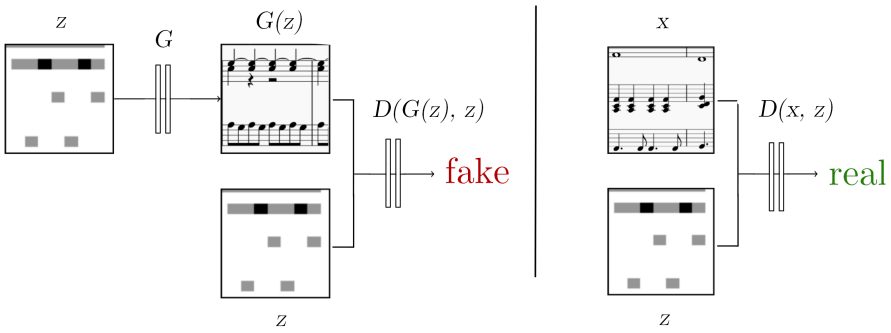

Fig. 2. ChordGAN architecture.

ChordGAN uses loss functions inspired by conditional GAN architecture. The generator loss is calculated by adding the L1 loss, or the mean squared error between the fake and real samples, to the 'fooling' likelihood: the probability the discriminator determines fake samples as real ($D_{real}(G(z), z)$). The L1 loss is multiplied by a constant λ for normalization purposes. During experimentation, we set λ to 100. The discriminator loss is comprised of the probabilities of the discriminator correctly identifying fake samples as fake ($D_{fake}(G(z), z)$), and real samples as real ($D_{real}(x, z)$)). These loss functions were inspired by Pix2Pix architecture (Isola et al., 2018).

$$G_{loss} = \mathbb{E}(D_{real}(G(z)), z) + \lambda \mathcal{L}_{L1}(G) \tag{1}$$

$$D_{loss} = \mathbb{E}(D_{real}(x, z) + \mathbb{E}(D_{fake}(G(z), z) \tag{2}$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|x - G(z)\|_1] \tag{3}$$

3 Implementation

3.1 Dataset

MIDI is a format for storing music note data that can be easily integrated with GAN training. Three datasets of MIDI files were utilized for testing purposes. The pop database consists of 122 snippets of contemporary pop songs. The second is comprised of various jazz excerpts (Malik & Ek, 2017). The third dataset is comprised of preludes of Johann Sebastian Bach, BWV 553–560. A single artist was chosen for the classical label to increase homogeneity; however, the pieces chosen from Bach's repertoire do not represent the genre in its entirety. Further artists from the Classical era were used with this model after experimentation, including excerpts from Mozart and Haydn. These trials produced similarly realistic results.

To increase the number of samples, each prelude, which had multiple musical sections, was split up into 16-measure intervals. The number of samples for each dataset was limited to 100. All three datasets can be found on the project website.[2]

3.2 Data Pre-Processing

Specific criteria were developed to standardize training data; for example, all pieces with time signatures outside of 4/4 were filtered out, as well as all pieces shorter than 50 timesteps. Additionally, while MIDI varies note velocities on a scale of 1 to 127, all note velocities were standardized to 127. Finally, all pieces were transposed to the key of C Major or a minor. Finally, to be read by the network properly, all MIDI files were converted to the piano roll format via the pretty-midi library (Raffel & Ellis, 2014). Further custom scripts were used to process the piano roll data, which included start and stop markers for notes. This conversion represented the MIDI file as a *numpy* array, with 16 time steps per bar.

3.3 Training

Training progressions for each genre tested are shown in Fig. 3. While the networks that trained on pop and classical data converged relatively early, at around 200 thousand epochs, the networks that trained on jazz converged late into training, at around 500 thousand epochs.

GANs experience a variety of difficulties during training (e.g. mode collapse, failure to converge) so additional steps were taken to improve transfer efficacy. For the final style transfer models, pop and classical training were paused for early stopping at observed convergence points, while jazz training was permitted to run for the full 500 thousand epoch duration.

[2] `https://conanlu.github.io/chordgan/datasets`

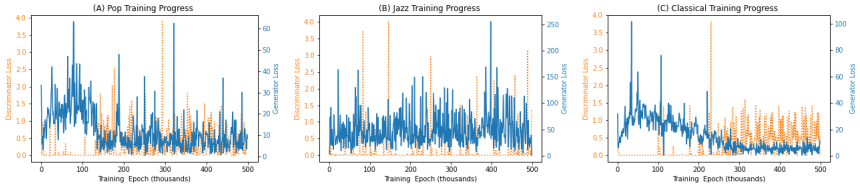

Fig. 3. Training progress graphs for each genre. The blue solid line tracks the generator loss, while the orange dotted line tracks the discriminator loss.

3.4 Demonstration

Transfer examples and code can be found on the project website.[3] One sample is shown below in Fig. 4.

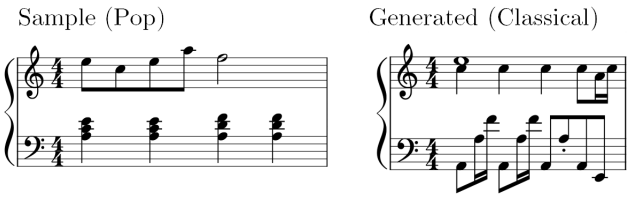

Fig. 4. Sample transfer. The measure on the left was extracted from a pop piece. It was given as input into the classical network, and generated the measure to the right. While harmonic structure remains constant within the measure, salient changes in note rendering are present in the classical style.

4 Evaluation Results

Once all three genre models were trained, several pieces from various genres were given as input to the network for conversion. 150 pieces of music outside of the original training datasets were utilized for testing purposes, with 50 in each genre. Two criteria were used to evaluate the success of the style transfer. Tonnetz measured the preservation of content, quantified in this paper as harmonic structure, while the independent genre classifier measured transfer realism.

4.1 Tonnetz Distance

The Tonnetz representation displays harmonic relationships within a piece (Wang & Dubnov, 2014). Since the main goal of style transfer, in this method, is to

[3] `https://conanlu.github.io/chordgan/samples`

retain the main harmonies and chords of a piece while changing stylistic elements, the Tonnetz distance provides a useful metric in determining the success of the transfer. A more similar Tonnetz graph between the post-transfer and pre-transfer pieces corresponds with a lower distance. This indicates the success of the transfer in maintaining content in terms of chords and harmony.

The Tonnetz of an individual piece of music is calculated by computing the dot product of the base Tonnetz graph with the chromagram representation of the piece of music. For the base Tonnetz graph, various intervals are calculated by iterating through the twelve different tonal categories with dimensions {6, 12}, as shown in Table 1. The Tonnetz matrices of pieces have dimensions {6, B}, where B signifies the number of bars.

Table 1. Base Tonnetz Graph

Fifth X	Fifth Y	Minor Third X	Minor Third Y	Major Third X	Major Third Y
$sin(\frac{7\pi}{6}x)$	$cos(\frac{7\pi}{6}x)$	$sin(\frac{3\pi}{2}x)$	$cos(\frac{3\pi}{2}x)$	$\frac{1}{2}sin(\frac{2\pi}{3}x)$	$\frac{1}{2}cos(\frac{2\pi}{3}x)$

The Tonnetz distance d between two pieces is calculated as the $L2$ norm of one Tonnetz matrix subtracted by the other. The subtracted matrix is denoted by a.

$$d = [\sum_{i,j} abs(a_{i,j})^2]^{1/2} \tag{4}$$

In this evaluation, the distance was calculated and averaged over the 16 measures of each sample. Each genre used 50 samples of 16-measure snippets, comparing a post-transfer piece to their pre-transfer counterparts. These formed the experimental tests. To validate the experimental results, the same 50 post-transfer pieces were compared to random pieces in a separate dataset, calculating the Tonnetz distance. The average Tonnetz distance within experimental pairing sets in all genres was found to be 0.00. In comparison, the average Tonnetz distances within control pairing sets were all measured to be higher than 0.00 (Table 2). Thus, based on the Tonnetz distance, it can be concluded the chords used remained constant throughout the transfer, indicating a continuity of content.

4.2 Genre Classifier

The second metric for evaluating the success of the genre transfer was an independent genre classifier. The model was a convolutional neural network (CNN) inspired by previous classification efforts (L. Roberts, 2020). In the original implementation, various spectral features were extracted. Within this evaluation, the convolutional neural network uses these features to categorize audio files into pop, jazz, and classical genres. Since the network used the *.wav* format, all MIDI files were converted to *.wav* with soundfonts. The genre classifier had an overall accuracy of 80% pre-evaluation. The evaluation consisted of 50 post-transfer

MIDI files from each genre given as input to the pre-trained genre classifier network (Table 3). These accuracy rates are higher or comparable to past implementations of symbolic music style transfer. However, the definition of style transfer in this paper's context resulted in more dramatic changes in musical composition overall, making such comparisons inapplicable. Future work to evaluate transfer quality may include subjective studies or surveys to gauge human perspectives.

Table 2. Tonnetz Distance

	Experimental	Control
Pop	0.000	0.044
Jazz	0.000	0.048
Classical	0.000	0.049

Table 3. Genre Classifier Accuracy

Baseline	80%
Pop	68%
Jazz	74%
Classical	64%

5 Conclusions and Future Work

ChordGAN achieves symbolic music style transfer using conditional GAN architecture. Chroma feature extraction was incorporated within the training process as a way to preserve content feature, such as the underlying chords and harmonies, while distinguishing the style involved in rendering chords to notes. Two metrics were used to evaluate the success. ChordGAN, through its success in both maintaining content while changing style in evaluations, can be considered an effective style transfer network. Current limitations for ChordGAN as a method for style transfer include the lack of long-term structure captured and training instability for certain genres like jazz. More sophisticated GANs like CycleGAN (Brunner, Wang, et al., 2018) have the opportunity to yield more realistic results. In particular, using chroma feature extraction in conjunction with known human-created composition rules has the possibility of yielding more realistic transfer outputs. The high efficacy of transfers makes ChordGAN a useful compositional tool. First, work is being done to map chords and lead sheets to the chromagram representation format, making automatic music generation with the network a possibility. The style elements of various compositional techniques can be synthesized, learned, and transferred, contributing to the composition process.

Acknowledgements

Creating ChordGAN would not have been possible without the help of teammates Elena Atluri and Satvik Nagpal, who contributed to the exploratory project GANmidi at COSMOS 2019.

References

Argamon, S., Burns, K., & Dubnov, S. (2010). *The structure of style: Algorithmic approaches to understanding manner and meaning.* Springer Publishing Company, Incorporated. doi: 10.5555/1869899

Brunner, G., Konrad, A., Wang, Y., & Wattenhofer, R. (2018). *Midi-vae: Modeling dynamics and instrumentation of music with applications to style transfer.*

Brunner, G., Wang, Y., Wattenhofer, R., & Zhao, S. (2018). *Symbolic music genre transfer with cyclegan.*

Cambouropoulos, E. (2016). The harmonic musical surface and two novel chord representation schemes. In D. Meredith (Ed.), *Computational music analysis* (pp. 31–56). Cham: Springer International Publishing. Retrieved from `https://doi.org/10.1007/978-3-319-25931-4_2` doi: 10.1007/978-3-319-25931-4_2

Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2017). *Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment.*

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). *Generative adversarial networks.*

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2018). *Image-to-image translation with conditional adversarial networks.*

Malik, I., & Ek, C. H. (2017). *Neural translation of musical style.*

Raffel, C., & Ellis, D. P. (2014). Intuitive analysis, creation and manipulation of midi data with pretty_midi. In *Proceedings of the 15th international conference on music information retrieval late breaking and demo papers* (Vol. 15).

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2019). *A hierarchical latent vector model for learning long-term structure in music.*

Roberts, L. (2020). Music genre classification with convolutional neural networks. *Towards Data Science*.

Wang, C.-i., & Dubnov, S. (2014, 10). Guided music synthesis with variable markov oracle.. doi: 10.13140/2.1.2171.2329

Yang, L.-C., Chou, S.-Y., & Yang, Y.-H. (2017). *Midinet: A convolutional generative adversarial network for symbolic-domain music generation.*