Arranging an Audio Track to other Genres by using CycleGAN-based Deep Learning Model *

Alex DongHyeon Seo
dseo22@wisc.edu

Hyecheol Jang
hyecheol.jang@wisc.edu

Stella Kim
ykim736@wisc.edu

1. Introduction

In this project, our goal is to change the genre of a music track, given a set of user inputs containing song and the desired genre. To accomplish the goal, the main model that we are going to consider is CycleGAN [6]-based model. Though Isola et al. proposed this architecture for unpaired Image-to-Image translation, we are hoping that this model with a proper modification of the structure would accept relative information from audio tracks.

1.1. Related Works

conditional GAN Image-to-Image Translation Translating a type of image to another type of image can be viewed as translating pixel by pixel of an image. Thus, instead of having different algorithms for different types of image translation [5], uses conditional Generative adversarial network(cGAN) to provide general solutions to any type of image translation. For any type of image translation, the same architecture and objective are used but is trained using different paired image dataset through Conditional GAN by putting conditions on the input image and outputting the corresponding image.

CycleGAN for Image-to-Image Translation Often finding a large size of paired dataset can be very expensive and difficult. On the other hand, unpaired dataset can be a little more reasonable to find compared to the paired dataset. Thus, Zhu et al. [6] provides a solution to such problem where the training samples do not have to be paired. Zhu et al. [6] did image-to-image translation which is translating a type of image to a different type using CycleGAN. CycleGAN's approach is that, instead of learning using paired dataset for training, learning from unpaired dataset of source domain X and target domain Y. Its strategy is that given domain X, the translated version of X to domain Y type should be translated back to domain X type without much loss.

Since it is hard to find paired dataset for arranged music, cycleGAN's approach is more suitable for our research, compared to cGAN.

CycleGAN in Music CycleGAN can not only learn from image dataset, but also can be used to learn from different types of dataset. One example is using CycleGAN to learn from MIDI files to do music genre transfer [2]. This example takes advantage of the concept of domain translation of CycleGAN and applies the concept to MIDI files with an extra discriminator to maintain the original structure of the input song.

Even though the purpose of this research is very similar to what we are trying to achieve in our own project, as this research utilizes MIDI files while our project uses audio file(MP3)s as input, the way to approach the problem must be different. Therefore, the uniqueness of the topic still exists in our project.

Music Genre Classification To clarify the genre of the given music, Bahuleyan [1] proposed a model using CNNs with spectrogram of given sound clip. The spectrogram is one of the ways to represent the sound data as an image, and various researchers has confirmed that treating soundwave as image by using spectrogram for machine-learning purposes works. As his dataset only contains 10 seconds of audio extraction from YouTube, he can have fixed size of image for input.

Outputs of this research can be used during the evaluation period. We might put our generated arranged music as an input of this research's model and check whether this model is able to classify our generated music as expected.

Since this research only utilizes a portion of music, while we need to analyze the entire sound track and generate similar length arranged music, we still need to find a better way to pass information about soundtracks to our model.

2. Motivation

Recent advancements of Artificial Intelligence grant computers the abilities to mimic the noble creativity of

*Project proposal for Spring 2020, University of Wisconsin-Madison STAT453 Deep Learning course (Instructor: Sebastian Raschka);
All authors contributed equally

the most intellectual lives. One of the advancements is computer-vision which allows computers to understand images just like human beings. However, unlike images, music has not been studied as extensive in the field of deep learning. Thus, we thought it would be interesting to do a project using audio files.

To choose a unique topic for our project, we decided to work with a model that generates music. We are inspired by Amazon Web Service's DeepComposer [3], a keyboard that allows people to create a new song or arrange to a different genre with a simple combination of notes of their own. DeepComposer is based on deep learning models using generator and discriminator to update the music. Slightly different from DeepComposer, our project will focus on transferring the genre of the song to the user's desired genre.

3. Evaluation

It is hard to find a great metric to measure how good the generated music is, since there is no numerical metric for evaluating the quality of music. As we are aiming to generate an arranged music while minizing awkwardness of sound itself and maximizing the similarity with other music of the targeted genre, we are required to set a proper way to numerically illustrate the performance of our model.

Though we need to investigate the best solution for this complicated problem, we have two possible ways to show the performance of our model. The first way of evaluating our model performance is to utilize different machine-learning models to classify the song's genre. If our generated music is classified to the targeted genre, we could mark it as a successful case.

The other possible way to check the performance of our model is to conduct a poll. With well-organized poll, like Isola et al.'s work [5], we are expected to get evaluation on how well our model can minimize the awkwardness of the generated song by asking people "awkward vs appropriate" after providing them a short clip of the generated music.

4. Resources

4.1. Datasets

Finding a large amount of music having paired labels(the arranged/mixed music based on the same track) is very challenging and even considered as impossible task, since there is a limited number of arranged tracks for each music(with high possibility, there is no arranged track). To use cycleGAN [6], we at least need to have unpaired labeled data. More specifically, we need music sources which have genre information as their label. Also, because the music industry is one of the most sensitive markets toward copyright issue, it is also important to only use the music tracks which do not have any restrictions on the usage for research purposes.

A good database that has music tracks that meet all the requirements illustrated above is Free Music Archive(FMA) [4]. It offers 106,574 untrimmed tracks of 161 genres, 879 GiB of data in total. FMA provides four different sizes of MP3-encoded audio data: *small, medium, large, and full*. Except for the *full* dataset, they only have trimmed (30 second length) tracks. Since we want to convert the entire music style, we are planning to retrive the list of music from *small* dataset and retrieve the actual music track from *full* the dataset. By doing so, we are hoping to limit the number of training data so that we can train our model more efficiently.

4.2. Computer Hardware and Computational Tools

To use GPU-accelerated model training, we will be using one of Google Colab, Center for High Throughput Computing (CHTC) of our university, or our own laptop that has NVIDIA GeForce 960M. Google Colab is the easiest tool to collaborate while CHTC is expected to provide the best computing power among these three options. Working locally will be the least preferred option, because of its limited performance and there could be other programs or softwares that are running concurrently with our training model. We will mainly use Python and PyTorch for our computational tools with Jupyter Notebook and PyCharm, the Integrated Development Environment, to develop the model.

5. Contributions

Our work will be divided fairly and evenly to make every group member to participate in both experiments and writings. We will divide the workload into three sections, 'Data engineering', 'Modeling' and 'Testing'. All three group members will contribute to all sections, but we will have a section leader for each section. Alex Seo will be leading the 'Data engineering' section, Hyecheol Jang will be leading the 'Modeling' section and Stella Kim will be leading the 'Testing' section. By doing this, every team member will actively contribute to all of the three sections and also has the responsibility to lead and complete their assigned section. The group leader for each section will also be responsible for the writings for the section. Writings for 'Abstract', 'Introduction' and 'Related work' will be completed together.

References

[1] H. Bahuleyan. Music genre classification using machine learning techniques, 2018.

[2] G. Brunner, Y. Wang, R. Wattenhofer, and S. Zhao. Symbolic music genre transfer with cyclegan. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 786–793. IEEE, 2018.

[3] D. Deahl. Amazon's new "ai keyboard" is confusing everyone, Dec 2019. `https://www.theverge.com/`

2019/12/4/20994203/amazon-web-services-deepcomposer-ai-keyboard-confusing-everyone.

[4] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016. GitHub Repository: https://github.com/mdeff/fma.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

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