Comparative Analysis of Music Style Transfer using Deep learning Techniques

Jainil Viren Parikh, Shreya Sahay, Karthik K S, Abhishek R, Amisha

Department of Computer Science & Engineering

R V College of Engineering,
Bangalore, India

jainilvirenp.cs17@rvce.edu.in , shreyasahay.cs17@rvce.edu.in, karthikks.cs17@rvce.edu.in
abhishekr.cs17@rvce.edu.in, amisha.cs17@rvce.edu.in

***Abstract*— Observing the success of neural nets in other fields this paper aims to produce interesting results from music . Music is fundamentally a sequence of notes. A composer constructs long sequences of notes which are then performed through an instrument to produce music. Humans can easily identify the genre of a music by just listening to it .However it is quite difficult to parametrize the style of music as it is not dependent on any fixed metric such as pitch,vocal etc but depends on its composition and performance which need to be correlated. Although the genre of music (Classical , Jazz etc)depends on the listener, there are certain distinguishing characteristics of each genre .Generative models can be applied to change properties of existing data in a principled way, even transfer properties between data samples.In this report we will aim to understand the styles of the different music genres and transform one style of music to another through transfer of properties. This is a comparison between two approaches and the resulting musical pieces, namely, Autoencoder-Generative Adversarial Networks (VAE-GAN) and Spectrogram Analysis methods**

***Keywords*— VAE-GAN, Spectrogram, Style Transfer**

I. Introduction

Crafting music can be traced back to a rule based vocal-to-pitch mapping algorithm developed by Guido D'Arezzo a famous musician whose algorithm generated a sequence of notes.The procedure of crafting music or algorithmic composition has become popular these days due to the sudden increase in CPU powers which helped researchers to try newer approaches quickly. Music style Transfer which is a part of algorithmic composition, focuses specifically on changing the characteristics of a song such that it sounds as if it were sung by another musician. Companies like Facebook have successfully been able to classify music i.e to differentiate music of one genre from the other ,we aim to recreate music of a specific genre by understanding the characteristics of different genres. Recent breakthroughs in deep generative models has led to the use of GAN's in automotive music generation. In this paper we explore two different techniques both employ deep neural networks to understand the intricacies involved in musical notes. Despite these promising methods it is to be noted that both natural and creative music generation is not feasible as an algorithm with weak constraints is often too random and rarely generates human-like music and an algorithm with strong constraints is too flat and they lack the dynamic that can be felt in naturally created music.

This paper discusses two approaches to music style transfer and offers a comparison between the two, carefully bringing out the use cases that each of these techniques excel in.

II. Literature Review

The idea of interpreting and building an understanding of the artistic styles of various pieces of art has been an area of interest since the beginning of 2010s. This has especially been explored and developed in the image domain.

The ground breaking work in this aspect [1] looks at decomposing an image into its style and content and recombining them to produce a result. This decomposition and recombination paved the way to understanding how deep neural networks can be made to learn artistic features and obtain their neural representations in spatial domain. Higher layers of a deep convolutional neural network that is trained for object recognition essentially capture the content information and the texture feature space contributes to the style information. Following this, an attempt was made at providing a style migration model [2] that improved on the per pixel losses as used in [1].

Once the use of CNNs in neural art style transfer reached a sense of maturity in terms of the research carried out, its potential applications caught the eyes of music scientists who were curious to see whether and how the highly capable neural networks can understand the artistic styles in music in addition to the sequence of notes and pitches.

The very first stage in this direction was the birth of music genre classification networks that could derive some latent representation for the style of music which is not mathematically parameterizable. The only way to characterize a music genre is by the common features shared by its members. The paper [3] discusses the classification of music into genre hierarchies utilizing both whole-frame as

well as real time scenarios using beat histograms to represent the styles. This paper [4] uses harmonic structures like chords to classify Brazilian music. It establishes a random forest model for classification into nine genres. Similarly, multiple approaches that optimize the genre classification have been proposed and are seen in many commonly used applications like Spotify, SoundCloud, shazam.

With a proper genre classification system in place, research interest progressed towards generating music and further control the style of music being generated.

The paper [5] proposes a scientifically-viable definition of music style transfer by breaking it down into precise concepts of timbre style transfer, performance style transfer and composition style transfer. This paper [6] proposes a variational autoencoder setup to handle music with multiple music tracks to change the pitch, dynamics and instruments in order to convert an input music from one style to another, [7] developed a system based on a WaveNet autoencoder [8] that can translate music across instruments, genres and styles, and even create music from whistling. The use of autoencoders, as in [9] based on a recurrent highway gated network combined with a variational autoencoder with filtering heuristics allows generating pseudo-live acoustically pleasing and melodically diverse music.

III. Methodology

This paper explores the task of music style transfer in two approaches, namely, spectrogram analysis and VAE-GAN(Variational Autoencoder- Generative Adversarial Network).

A. VAE-GAN :

We propose a VAE-GAN architecture(Variational Auto-encoder Generative Adversarial Network). By combining a variational auto-encoder with a generative adversarial network , the learned feature representations in the GAN discriminator can be used as the basis for the VAE reconstruction objective. This makes the training process significantly more stable as opposed to using just a VAE like in [9] as the generator has information regarding the real-world entities it is trying to generate from the discriminator of the GAN rather than guessing what the real-world entity should be at each iteration. In addition, the Encoder learns the mapping of images to Latent space which is very useful.

VAE architecture:

The variational autoencoder is built as given in fig 1. The encoder is composed of conv2D, batch normalization and dropout layers that take as input a batch of images of size 88x88 and produces an intermediate representation which acts as an input to the decoder. The decoder uses a series of conv2D transpose layers. The end result of this autoencoder is the regenerated music representation based on the knowledge of genre features or characteristics that the autoencoder gained during the training phase.

GAN architecture:

GAN architecture makes use of two components,namely, a generator and a discriminator. The generative model is pitted against an adversary, which is a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution [10]. In our work, the decoder from the variational autoencoder described above and as depicted in fig 1 plays the part of generator for the GAN. The music generated by the decoder is then passed through a discriminator in order to determine whether or not the produced resultant music is as per the required genre characteristics. The discriminator used is composed of conv2D and dropout layers. Since the generator tries to produce music such that the discriminator may classify as that of the required genre and the discriminator tries to improve its ability to decide whether an input to it was generated by the generator or it was an actual music from the required genre, they supplement each other in increasing the model accuracy to produce music in the target genre with the content from the source genre.

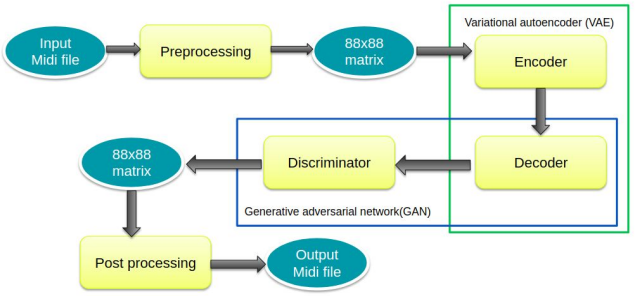


Fig. 1 VAE-GAN architecture

Dataset preprocessing:

The dataset was taken from [11] which is distributed under CC-BY-4.0 license. It consists of 200 piano only songs in midi format classified into jazz and classical with an average length of 4 minutes each. For generating input to the model, each midi file was aligned to a specific time interval and converted into a 88x88 size matrix. This matrix representation essentially indicates that 88 time steps were taken along 88 pitches that are possible in a piano. Each entry in the matrix represents the velocity of the particular note at every timestep.

Training :

The training of both VAE and GAN was done simultaneously. Therefore, multiple losses incurred at various stages of this pipeline were considered for optimization using an Adam optimizer at each stage as in fig 5.

The variational autoencoder loss function was approximated as a negative sum of the reconstruction error and a prior term for regularization.

$$loss_{vae} = -E_{q(z|x)} log\left[\frac{p(x|z)p(z)}{q(z|x)}\right]$$

The reconstruction error calculated here is the cross entropy reconstruction loss as :

The GAN consists of two networks: the generator network Gen(z) maps latents z to data space while the discriminator network assigns probability y = Dis(x) ∈ [0, 1] that x is an actual training sample and probability 1 − y that x is generated by the model through x = Gen(z) with z ∼ p(z).

The losses for the GAN are calculated below where both the generator and discriminator use sigmoid cross entropy to calculate their individual losses. The GAN loss is a combination of both these losses incurred.

$$-(tensor_{target}) * log(tensor_{output}) - (1 - tensor_{target}) * log(1 - tensor_{output})$$

A very important role played in the quality of training and speed of convergence is the use of dynamic learning rates for both the generator as well as discriminator. With each epoch, the learning rate is modified based on the losses incurred in the previous epoch.

The resultant output is a 88x88 array that needs to be converted back into its midi representation using the pitch and volume information encoded in the matrix at various timesteps. This post processing completes the procedure for VAE-GAN based music style transfer

B. ***Spectrogram analysis:s***

In the proposed spectrogram analysis method two audio files are taken as an input. This method follows three phases :The audio files are converted to their respective spectrogram images , The spectrogram images are analyzed and the result spectrogram image is obtained,an audio file is reconstructed from the resultant spectrogram image.

Spectrogram Preprocessing :

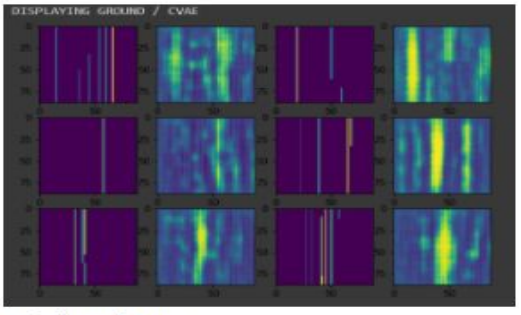
A spectrogram is a 2 D representation of the frequencies of a 1D signal as it varies with time.SPectrograms are generally viewed as a 1xT image of F channels. Spectrograms finds its use in fields such as music automation , sonar and speech processing . To convert an audio file to a spectrogram image we use short time fourier transforms. We apply fourier transforms of the 1D audio signal and convert them spatial domain to frequency domain. To enable ease of computing we use Fast Fourier transform which is a computational faster implementation of fourier transform. Once the spectrogram images are formed we feed them to the analysis stage.

Spectrogram image analysis:

We propose a spectrogram analysis method which combines the information present in a spectrogram of a 1D signal with the famous style transfer method proposed by Leon Gatys[1] et al. for images.The success of neural art transfer for image related tasks has been to start with two images a content image and a style image and then optimize the input signals starting with random noise to take the features of interest from both the content and style images. Here we employ a similar procedure where we take the spectrogram images of two audio signals and starting with a random noise we transform it to a combination of the two input audio signals.To extract features from the spectrogram images we require a neural network that can understand the complex patterns of the spectrogram image. We first experimented with VGG-19 a state of the art Image classification network that was trained on the imagenet dataset. Being trained on image datasets we assume that the network will easily be able to extract important features from the image. But as spectrogram images are not general RGB images and 3x3 convolutions are not well suited for our 1D problem. So we move to a second method that uses 1D convolutions. Here we can either use a pretrained network that uses 1d Convolution or a network whose weights have been randomly initialized. Analysis of both the approaches shows that the random weights implementation has the same effects as the pretrained models. So we further move with the randomly initialized weights approach for image feature extraction. The features are extracted from both the input audio file spectrograms and they are compared with the features from the random noise. The loss is calculated from the loss functions shown in below. Then using L-BFGS we update the weights(random noise vector) such that the loss is minimized. After the algorithm has converged the resultant spectrogram image now represents the combination of the input spectrograms.

$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a}, \vec{x})$$

α and β are the weighting factors for content and style reconstruction



Training of GAN

Output:

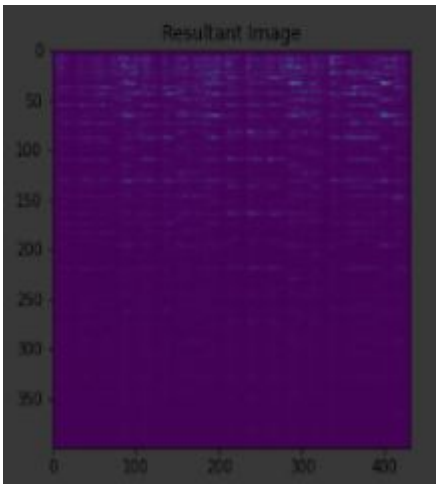
The resultant spectrogram image needs to be converted from frequency domain to a 1D signal along with phase information. The 1D signal is obtained by Griffin-Lim algorithm .This algorithm is computationally simple and is obtained by minimizing the mean squared error of the STFT(short time Fourier Transform) of the estimated signal and the modified STFT. The resultant is a phase regenerated audio file which is a mixture of both the input audio files.

C. Results and observations :

Spectrogram Results:

We evaluated the proposed spectrogram analysis approach using a music genre classifier that was trained on GTZAN . Figure 3 shows the different spectrogram images that were resultant from the VGG model and the 1D convolution models.As can be observed the output from the VGG net is extremely noisy.This is mainly due to the fact that VGG is trained for image description . The shallow net architecture with randomly assigned filters provides the best result. To better understand the results we passed the input audio files and the resultant audio file to the music genre classifier. What we observed is that the output obtained from the VGG model was classified to a genre which is neither of the input audio genres.This clearly proves that the resultant from VGG is noisy. We apply the same classifier to the resultant audio file of the shallow network. What we observed is that the output genre is a mixture of the input genres which is the expected result.One can also hear the outputs from the shallow networks which seem to place the style texture most precisely . Our tests confirm that the use of random shallow networks with L-BFGS has provided smoother results compared to gradient descent approach.

Fig 3: Random shallow net output:



Gan Outputs:

The proposed VAE-GAN model was evaluated by passing the input and output music through a music classifier and the change in genre was clearly noticed in favour of the target genre. However, the resultant music files were found to be highly noisy as in fig 6, and it can be said that the style transfer occurred at the expense of conserving content information. This ill effect however can be overcome by training over more number of epochs and better control over the dynamic learning rate. The noise observed in the model outputs also can be credited to the fact that the only information considered for style understanding was pitch and volume at each timestep. Other composition aspects were not taken into account.

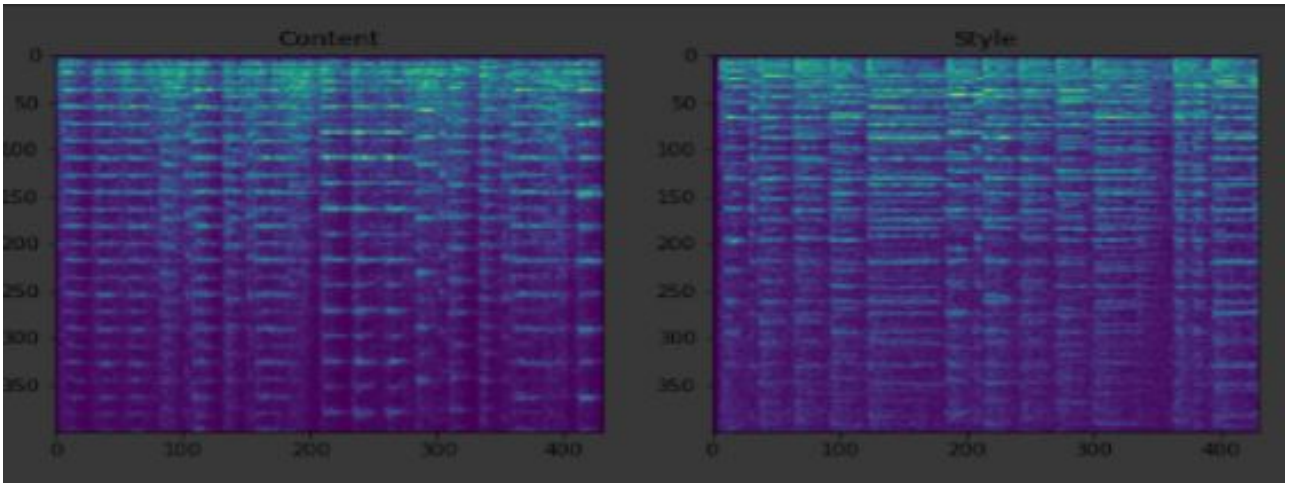
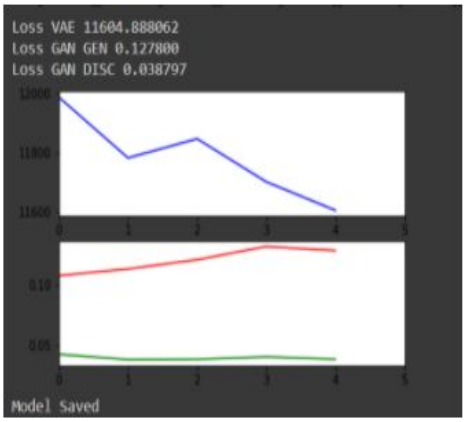


Fig 4:Input content and style music Spectrogram Images



loss graphs during training

Fig 5 : Loss graphs during GAN training

In view of this, the following approach to enhance performance of the VAE-GAN model is proposed. The music midi file representation into the matrix format here only captures the velocity of every note at timesteps. However, efforts can be made to encode more relevant information in the matrix concerning the pitch transitions and melody. In addition to this, our model was trained at random pieces of music files of 88 steps each. This could have resulted in scenarios where the number of notes being played are too less or is filled with silence which in turn contributed to the performance degradation.

*D. **Conclusion:***

In this work we proposed two methods for audio style transfer. One that used generative adversarial networks to understand the intricacies of single track piano music, and the spectrogram analysis model that interpreted 1D signals in frequency domain over multitrack music as well. The experiments showed that the resultant audio files from the GAN model were noisier than the spectrogram files but showed prospects for a much greater improvement in quality over a slightly more fine-tuned training from the aspect of learning rates than the spectrogram results. By comparing the results of both the models with a music genre classifier we come to a conclusion that generative models outperform spectrogram models. However, spectrogram analysis provides us with an interesting insight into the image representations of audio files. Our research shows that spectrogram analysis has also paved the way for using pre trained networks which were meant for image related tasks to be used on audio files. This research topic is very unique and at this stage numerous other relevant sound texture models should be further explored in future work.

*E. **References***

[1] Gatys, Leon A., et al. “A Neural Algorithm of Artistic Style.” ArXiv:1508.06576 [Cs, q-Bio], Sept. 2015. arXiv.org, http://arxiv.org/abs/1508.06576.

[2] Y. Li, T. Zhang, X. Han and Y. Qi, "Image Style Transfer in Deep Learning Networks," 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, 2018, pp. 660-664, doi: 10.1109/ICSAI.2018.8599501.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, July 2002, doi: 10.1109/TSA.2002.800560.

[4] Wundervald, Bruna D., and Walmes M. Zeviani. “Machine Learning and Chord Based Feature Engineering for Genre Prediction in Popular Brazilian Music.” ArXiv:1902.03283 [Cs, Eess, Stat], Feb. 2019. arXiv.org, http://arxiv.org/abs/1902.03283.

[5] Dai, Shuqi, et al. “Music Style Transfer: A Position Paper.” ArXiv:1803.06841 [Cs, Eess], July 2018. arXiv.org, http://arxiv.org/abs/1803.06841.

[6] Brunner, Gino, et al. “MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer.” ArXiv:1809.07600 [Cs, Eess, Stat], Sept. 2018. arXiv.org, http://arxiv.org/abs/1809.07600.

[7] Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. A universal music translation network. CoRR, abs/1805.07848, 2018.

[8] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pages 1068–1077, 2017.

[9] Yamshchikov, Ivan P., and Alexey Tikhonov. “Music Generation with Variational Recurrent Autoencoder Supported by History.” ArXiv:1705.05458 [Cs], Nov. 2018. arXiv.org, http://arxiv.org/abs/1705.05458.

[10] Goodfellow, Ian J., et al. “Generative Adversarial Networks.” ArXiv:1406.2661 [Cs, Stat], 1, June 2014. arXiv.org, http://arxiv.org/abs/1406.2661.