Music Generation using Deep Generative Modelling

Advait Maduskar
Department of Information Technology
Thakur College of Engineering and Technology,
Mumbai, India
advaitmaduskar.1@gmail.com

Aniket Ladukar
Department of Information Technology
Thakur College of Engineering and Technology
Mumbai, India
ladukaraniket@gmail.com

Shubhankar Gore
Department of Information Technology
Thakur College of Engineering and Technology
Mumbai, India
shubhankar27gore@gmail.com

Neha Patwari
Department of Information Technology
Thakur College of Engineering and Technology
Mumbai, India
neha.patwari@thakureducation.org

2020 International Conference on Convergence to Digital World - Quo Vadis (ICCDW) | 978-1-7281-4635-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCDW45521.2020.9318683

Abstract - **Efficient synthesis of musical sequences is a challenging task from a machine learning perspective, as human perception is aware of the global context to shorter sequences as well of audio waveforms on a smaller scale. Autoregressive models such as WaveNet use iterative subsampling to generate short sequences that are a result of a localized modeling process but lacking in overall global structures. In juxtaposition, Generative Adversarial Networks (GANs) are effective for modeling globally coherent sequence structures, but struggle to generate localized sequences. Through this project, we aim to propose a system that combines the random subsampling approach of GANs with a recurrent autoregressive model. Such a model will help to model coherent musical structures effectively on both, global and local levels.**

Keywords – generative adversarial networks, autoregressive, music generation, deep learning

I. Introduction

Synthesis of audio is useful for a wide spectrum of applications including music generation systems. Audio generation algorithms, known as vocoders in TTS and synthesizers in music, respond to higher-level control signals to create fine-grained audio waveforms. Synthesizers have a long history of being hand-designed instruments, accepting control signals such as ‘pitch’, ‘velocity’, and filter parameters to shape the tone, timbre, and dynamics of a sound (Pinch et al., 2009). In spite of their limitations, or perhaps because of them, synthesizers have had a profound effect on the course of music and culture in the past half century (Punk, 2014) [1] .

Like any other form of information, musical sequences have underlying patterns that are indiscernible to human eyes. Humans compose new music based on experience and intuition and not mathematical patterns. Our objective is to replicate this kind of “thinking” in a computer and to that purpose, the deep learning approach can prove fruitful. We aim to train a generative network on existing musical data and learn patterns in the data in order to generate new sequences of medium to long length.

II. Literature Review

For the purpose of identifying shortcomings in the existing models for music generation, we performed a comprehensive literature survey. The papers we chose to study are recent papers with all of them published post 2016. The models we studied were largely based on deep generative modeling and consisted of LSTM (Shin A., Crestel L.), Transformer (Cheng-Zhi, et al.), AutoEncoders (Engel J., Resnick C., et al.) and GAN (Hao-Wen Dong, et al.) [2] [3].

Based on our literature survey, we identified shortcomings and opportunities in WaveNet and MuseGAN, the two models that produced the most significant results. WaveNet is a deep autoregressive model that processes data at a relatively fast rate. However, it struggles to model sequences on a global scope due to memory limitations of autoregression. Generative Adversarial Networks use random subsampling to achieve coherency at a global level and thus overcome the shortfall of the WaveNet. It uses two components – the generator and discriminator to generate subsequences and compare with source respectively. However, GANs fail when it comes to generating short term sequences due to overfitting on training data and are extremely expensive from a computational perspective.

This literature review has allowed us to propose the development of a model that combines the lightweight autoregressive approach of WaveNet with the random subsampling methodology of GANs to develop coherent medium-to-long term musical subsequences without having to conform to computing or memory limitations.

978-1-7281-4635-5/20/$31.00 ©2020 IEEE

III. BACKGROUND

One of the earliest papers on deep learning – generated music, written by Chen et al, generates one music with only one melody and no harmony. The lack of coherence and structure on a global perspective was one of the problems cited.

This suggests that there are two main directions to improve upon –

1) Generate music with musical cadence, more intricate structure, and utilizing all types of notes.

2) Produce a model capable of learning long- term structure and possessing the ability to build a descant and return to it in the section.

Liu et al tackle the same problem but are unable to overcome either challenge. They mention that their representation of music cannot accurately interpolate between the melody and the other parts of the piece, while also failing to note the complexity of musical pieces. [4]

Eck et al use two different LSTM networks - one to learn chord structure and local note structure and one to learn longer term dependencies in order to try to learn a tune and preserve it during the course of the piece. But this architecture is trained on a fixed number of chords and is thus unable to create a more diverse combination of notes. [5]

IV. SIGNIFICANCE

Music is a complex, highly structured sequential data modality. When rendered as an audio signal, this structure manifests itself at various timescales, ranging from milliseconds, all the way to a piece of music that is several minutes long. Modelling all of the temporal correlations in the sequence that arise from this structure is challenging, because they span many different orders of magnitude. [6]

The significance of this project is not limited to overcoming the challenge of music generation. It serves as a tool for taking human intelligence one step further by augmenting the ability of the composer to create music. Furthermore, the project will also act to detect plagiarism and act as a device to resolve copyright issues in the music industry. Finally, the project embodies and represents the effect of digitization and the cultural shift that comes with it, by integrating machine learning with one of the foundations of human culture – music. [7]

V. OBJECTIVES

Like any other form of information, musical sequences have underlying patterns that are indiscernible to human eyes. Humans compose new music based on experience and intuition and not mathematical patterns. Our objective is to replicate this kind of "thinking" in a computer and to that purpose, the deep learning approach can prove fruitful. We aim to train a generative network on existing musical data and learn patterns in the data in order to generate new sequences of medium to long length.

We are working towards hypothesizing a new model for music generation that combines the properties of models we previously studied while omitting the shortcomings and integrating the positive aspects. We aim to conceptualize a model that amalgamates the properties of AutoEncoders and GANs in a recurrent feedback loop.

VI. METHODOLOGY

We are focusing on classical music for training of the model because of the availability of MIDI files for classical music as well as classical music is relatively more standardized in contrast to other genres. The music21 library in Python will be used for processing of files.

Each song is to be converted into a string of chords, with each predefined time period corresponding to one chord. Every time a new note is played, a new chord will be created. Then, for processing, each chord will be represented as an encoded vector. A total of 700 songs will be used to give approximately 390,000 notes. The training-validation-testing split used will be a standard 70-15-15 split.

VII. PROPOSED SYSTEM

The concept of GAN is fundamentally unambiguous and simple. Both the Generator and Discriminator are neural networks competing against each other. The aim of Generator is to generate data which appears to be like the real data. The job of the Discriminator is to distinguish between real data and fake data generated by the Generator. Here we plan to train the Generator once for every five training iteration of the discriminator.

Autoencoders are a distinct type of feed forward neural networks. Their basic goal is to reduce the dimensionality of the input. Here, the input and output remain the same. Autoencoders consist of mainly an encoder, a dimensional code and a decoder [8].

`

The GAN based Autoregressive approach, which we have illustrated in Figure 1, can prove to be fruitful because GANs being very effective in modelling an underlying pattern will produce quality results. Meanwhile, Autoencoders can be used to compress the data without losing important features will aid in speeding up the training process. This will eventually result in a faster and better generation of music [9].

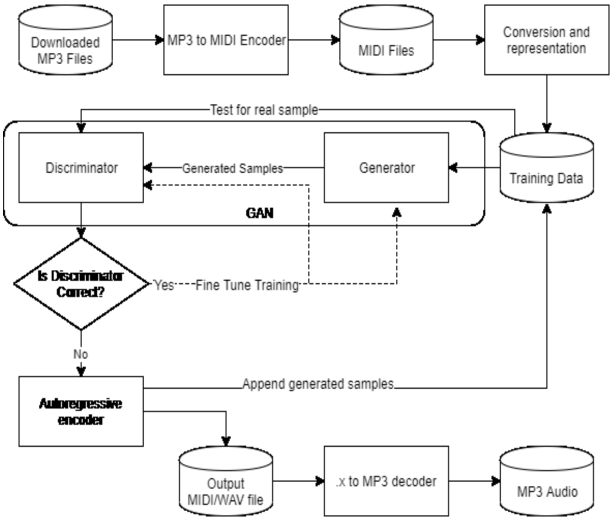

Fig. 1. The proposed system which combines a GAN based system with an autoregressive approach

VIII. Results Achieved

The following results, as seen in Figures 2,3 and 4, were achieved when a GAN based model was trained on a dataset consisting of Bach's musical symphonies. The conditional attribute used here was pitch. Using pitch, the generator learns to represent different instrument timbres using its latent space [10].

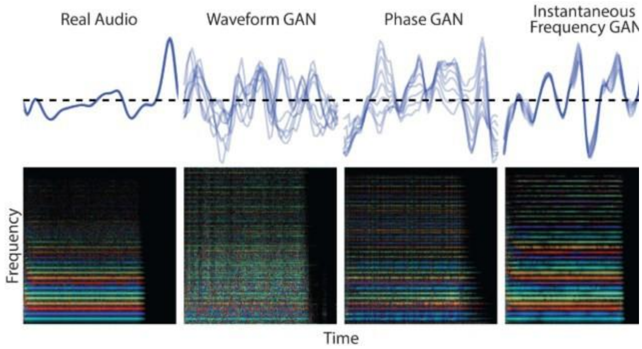

Fig. 2. Representation of musical waveforms

Here, we have used random interpolation to generate music on a scaled GAN architecture. After combining the MIDI files and parsing them, they can be represented as follows –

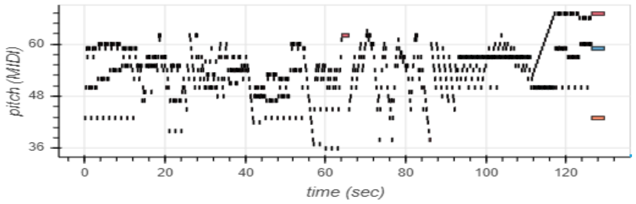

Fig. 3. Pitch vs Time representation of dataset used for random interpolation

Following the training of the model using this data, the time interval for random interpolation between instruments was set as 5 (sec/instrument) and constant Q-spectrogram of the generated music was obtained.

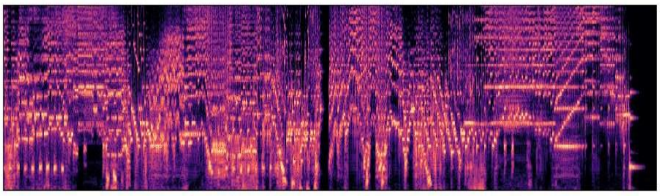

Fig. 4. Constant-Q spectrogram of generated music

Thus, from our dataset, we were able to randomly subsample, interpolate and generate 656 samples.

We observed that the interpolation between latent vectors was coarse and could be improved by increasing the time period between interpolations. Thus, to produce music that sounds smoother, we can set this time period to higher values and compare with currently produced music. Furthermore, more training data will help refine the model and help towards smoother conformation between instruments.

IX. Future Scope

In the near future, we are focusing on improving established baselines by manipulating training data and interpolation times. We also plan on conducting a comparative analysis with other methods for music generation with various permutation of functional parameters.

Future scope includes development of an autoregressive GAN model in a recurrent feedback loop that can be used not just for music generation, but also for detection of plagiarism in music by introducing original work as a sample data point and the dubious work as the fake samples in the discriminator function of the GAN.

`

X. Conclusion

Based off our model, we can extrapolate that a Generative Adversarial Network-based approach to music generation, while highly effective, is expensive in terms of memory as well as computation.

Combining Autoencoders with GANs and allowing the model to regress on itself can reduce the amount of data required while speeding up computation. This is possible due to the dimensionality reduction being done by autoencoders working with the sampling and generative abilities of GANs. This combined approach performs excellently on single-instrument music, since the autoregression makes for a uniformly coherent substructure formation.

References

[1] J. Engel, C. Resnick, A. Roberts, S. Dielman, M. Norouzi, D. Eck and K. Simonyan, “Neural audio synthesis of musical notes with wavenet autoencoders.,” in *34th International Conference on Machine Learning*, 2017.

[2] H. W. Dong, W. Y. Hsiao, L. C. Yang and Y. H. Yang, “MuseGAN demonstration of a convolutional GAN based model for generating multi-track piano-rolls.,” in *International Society of Music Information Retrieva Conference*, 2017.

[3] A. Shin and L. Crestel, “Melody Generation for Pop Music via word representation of musical properties,” *arXiv,* vol. abs/1710.11549, 2018.

[4] R. B. Liu I, “Bach in 2014: Music Composition with Recurrent Neural Network,” *arXiv Preprint,* vol. 1412, no. 3191, 2014.

[5] C. Z. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu and D. Eck, “Music transformer: Generating music with long-term structure,” 2018.

[6] S. K. D. S. Van Den Oord A., “The challenge of realistic music generation: modelling raw audio at scale,” *Advances in Neural Information Processing Systems,* pp. 7989-7999, 2018.

[7] G. H. F. P. JP Briot, “Deep learning techniques for music generation-a survey,” *arXiv Preprint,* vol. 1709, no. 1620, 2017.

[8] J. A. Henning, A. Umakantha and R. C. Williamson, “A classifying variational autoencoder with the application to polyphonic music generation,” *arXiv preprint arXiv:1711.07050,* 2017.

[9] I. P. Yamshchikov and A. Tikhonov, “Music generation with variational recurrent autoencoder supported by history,” arXiv preprint arXiv:1705.05458, 2017.

[10] J. Engel, K. K. Agarwal, S. Chen, I. Gulrajani, C. Donahue and A. Roberts, “Gansynth: Adversarial neural audio synthesis,” arXiv preprint arXiv:1902.08710, 2019.

`