Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

Music Generation using BERT-GAN

Aakash Goradia, Prachi Tawde

***Abstract*—The process of music creation exhibits several fundamental disparities when compared to the creation of images and movies. In order to effectively analyze and understand music, it is imperative to employ a temporal model, given that music is an art form that unfolds over time. Finally, it is worth noting that polyphonic music often organises notes into chords, arpeggios, or melodies, thereby rendering the introduction of a chronological sequence of notes unnatural. Therefore, it is important to use a generative model for sequential modeling. This paper presents a new method of generating symbolic music using BERT (Bidirectional Encoder Representations from Transformers) and Generative Adversarial Networks (GANs). The Lakh Pianoroll Dataset was used to train the model, which contains over 100,000 bars of rock music. It is shown that the model can produce original and coherent music from nothing.**

***Index Terms*—BERT, GANs, Transformers, Symbolic Multi-track Music Generation**

I. INTRODUCTION

It is hard to make music with the help of artificial intelligence (AI). It is hard to make original musical pieces that sound real. Rule-based systems [1] and statistical models [2] are two of the traditional methods to do this, but they don't fully understand the artistic and complex parts of music.

Generative Adversarial Networks (GANs) have become widely recognized in the last few years. It is a popular method for creating unique and real content. This content can include text, sounds, and pictures. A GAN has two main parts. One, the generator, makes new examples. Two, the discriminator, spots the differences between real and new examples. It then sorts them into separate groups. Recent research shows GANs can create attractive pictures and realistic sounds. This makes GANs a possible solution for composing and producing music. The creation of BERT (Bidirectional Encoder Representations from Transformers), has uplifted the grasp of language in natural language processing (NLP). Transformers, a type of structure, are used to build word-connections in BERT, a top-notch language model. The model got its title via lots of training with large sets of written data. This resulted in a high-level ability to sequence tasks. The rigorous training lets it understand intricate word relationships in the text.

The proposed methodology combines BERT and GANs for music generation because of GANs' proficiency in synthesising auditory signals and BERT's capabilities in natural language processing. In the context of the GAN architecture, the BERT model serves the dual role of a discriminator and a generator. This configuration can effectively distinguish between authentic and synthetic musical compositions by BERT while concurrently enabling the generation of music adhering to predetermined structures and regulations. The objective of this approach is to overcome the challenges associated with traditional song composition methods through the utilisation of GANs and BERT. The findings of our research contribute to the expanding corpus of research involving natural language processing, deep learning methodologies, and the domain of musical composition.

II. BACKGROUND WORK

A. Bert

BERT is a major breakthrough in NLP and has been widely used in various applications and tasks. It is a language representation model and a state-of-the-art system that performs exceptionally well. It possesses the ability to acquire profound bidirectional representations through the analysis of unannotated textual data. Every layer of the model incorporates data from both the previous and following contexts to achieve this. The use of contextual information has been instrumental in the development of state-of-the-art models for NLP tasks. The BERT model, originally developed for NLP tasks, has demonstrated its applicability beyond the realm of NLP. Specifically, it has exhibited its utility in the domain of music generation. The primary advantage of BERT resides in its capacity to comprehend contextual associations and intricate dependencies within natural language. The aforementioned distinctive capability possesses the potential to enhance the process of generating musical sequences, thereby propelling the field of music generation forward. The utilisation of BERT aligns

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

with the prevailing practice of incorporating deep learning techniques in music-related tasks, thereby offering a promising avenue for investigating the capabilities of language models in capturing complex patterns in musical data. The versatility and effectiveness of BERT in modelling relationships within sequential data are exemplified by its widespread adoption in both NLP tasks and music generation.

B. Gan

Generative Adversarial Networks (GANs), make lots of different stuff. A GAN introduced by Goodfellow et al. [3] has two parts. One part to make fake samples, called a Generator. Another to tell which sample is real or fake, known as a Discriminator. GANs have demonstrated remarkable outcomes in image synthesis by undergoing an adversarial training process. Their utility has also been extended to other domains, such as music generation.

The music generation using GANs is rising as the leading method for creating realistic musical outputs. Different types of music have been generated using GANs, like symbolic representations, audio renderings, and multi-track compositions. These models have been able to capture intricate patterns in music compositions as various architectures and training methods are explored that would improve the quality, diversity, and manipulation possibilities of the generated music. Machine learning methodologies have been integrated with GANs to enhance music generation.

III. LITERATURE REVIEW

GANs can be combined with the Transformer architecture [4] to improve the quality and controllability of music compositions. One notable study proposed a self-supervised task was devised to facilitate the transformation of user-provided instructions into proficient and manageable symbolic music. In a separate investigation, Transformer Variational AutoEncoder (VAE) models were utilised to generate music that is intelligible to human listeners. Utilizing a context-sensitive hierarchical representation, which facilitated the exploration of various musical progressions, allowed for this.

In a different study, Reinforcement Learning (RL) algorithms were used to build RL-Chord, a melody harmonization system generating high-quality chord progressions through a melody-conditional LSTM model. A novel approach involved training four Transformer-XL networks simultaneously on time-valued note sequences, resulting in improved music quality and length.

Muhamed et al. introduced a unique method for generating coherent, long-term music using adversarial training of Transformers. This research utilizes a pre-trained SpanBERT model as the discriminator and the Gumbel-SoftMax trick for differentiable sampling. The study introduces the Bar Transformer, a hierarchical model addressing challenges in long-term dependencies for generating musically coherent and structurally meaningful compositions.

Within the domain of pop music composition, an advancement has been made in the form of a Pop Music Transformer. This transformative technology has been created by leveraging the power of Transformer-XL and REMI, an innovative event representation derived from MIDI data. The present study was primarily concerned with the development of a methodology for the generation of highly expressive pop piano music that exhibits enhanced rhythmic structure. In this study, we present an alternative methodology known as Museformer, which utilizes a Transformer model to address the complexities associated with long sequence modelling and music structure modelling. To achieve this, we incorporate both fine-grained and coarse-grained attention mechanisms within the model. By doing so, we aim to effectively capture the intricate structures inherent in music compositions.

In the realm of music generation, Dong et al. introduced MuseGAN [5], an innovative approach that leverages the power of GANs. The primary objective of MuseGAN is to generate intricate musical compositions that are symbolic in nature, encompassing multiple tracks and exhibiting polyphonic characteristics. The utilisation of Convolutional Neural Networks (CNNs) was employed for both the Generator and Discriminator in this study. Moreover, this study presents the implementation of a Transformer-GAN model that is conditioned by human sentiment data. The primary objective of this model is to generate songs by leveraging the perceived valence and arousal associated with the input data. The present study endeavours to elucidate the intricate interplay between musical structure and affective experiences.

Researchers have already tried a variety of methods. For example, symbolic music data can be used to train diffusion models, continuous latent spaces of variational autoencoders that have already been trained can be used to show the discrete domain, and non-autoregressive generation can be used for parallel generation with a set number of refinement steps.

These studies collectively contribute to the progression of music generation techniques through the utilisation of GANs integrated with BERT architecture. They present novel methodologies for enhancing controllability, structural awareness, sentiment-based synthesis, and long-term coherence in the process of generating symbolic music.

IV. DATASET

A. Description

The Lakh Pianoroll Dataset (LPD) is an extensive aggregation of 174,154 multitrack pianorolls obtained from the Lakh MIDI Dataset (LMD). The Low Power Device (LPD) utilises a specialised format that has been specifically designed to maximise the efficiency of input and output operations and minimise storage consumption. The dataset comprises multiple subsets and versions, with a specific emphasis on LPD-5, the specific variant utilised in this research article. The LPD-5 device effectively consolidates the discrete tracks within a given dataset into five discernible classifications: Drums, Piano, Guitar, Bass, and Strings Fig.1. This categorization process is primarily reliant on the identification of programme numbers present within the corresponding MIDI files. In this study, we employ the lpd-5-cleansed subset, which comprises

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

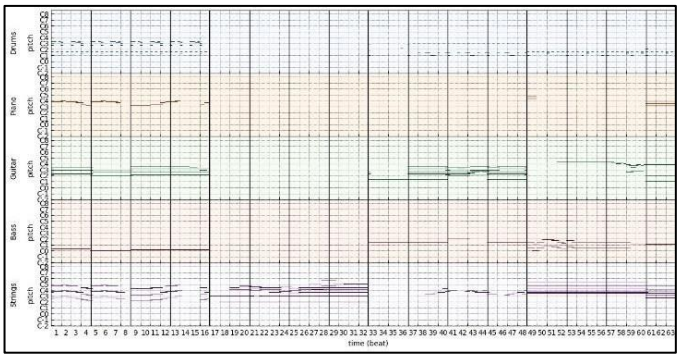

Fig. 1. Sample Dataset

a collection of 21,425 five-track pianorolls extracted from the larger lpd-cleansed dataset.

B. Preparation

Firstly, the multitrack data was loaded as a pypianoroll. Multitrack instance. Then, the pianorolls were binarized to convert them into binary format. Subsequently, the pianorolls were down sampled to a shape of 4 x 72, representing the desired number of timesteps and pitches, respectively

The pianorolls were then stacked in a shape of 5 x 4 x 72, aligning with the number of tracks, timesteps, and pitches, to facilitate further analysis. The relevant pitch range was extracted, ensuring that only pitches within the range of MIDI note number 24 to MIDI note number 95 were retained.

Next, the total number of measures in each sample was calculated based on the beat resolution and tempo. Random selection of a desired number of phrases was performed from the multitrack pianoroll. However, samples with insufficient notes in one or more tracks were excluded from the dataset to maintain data quality.

Finally, all the collected pianoroll segments were stacked together into a single comprehensive array, consolidating the dataset for further processing and analysis. Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

V. PROPOSED MODEL

The Bidirectional Encoder Representations from Transformers (BERT) model is employed in both the Generator and Discriminator components of the Generative Adversarial Network (GAN) architecture. The primary aim of the Generator is to produce sequences that exhibit similarity to those found in the authentic dataset without relying on any contextual information from the actual dataset. The intention is for these generated sequences to deceive the Discriminator Fig.2.

The Discriminator module is responsible for evaluating the authenticity of each input sequence by considering it as a unified entity. Its primary objective is to discern whether a given sequence is genuine or artificially generated. Drawing inspiration from [6], we divide the input sequences into patches of predetermined length and subsequently convert each patch into a singular feature vector.

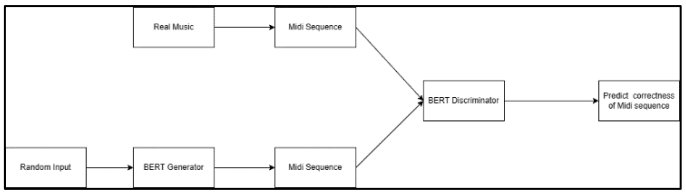

Fig. 2. Architecture

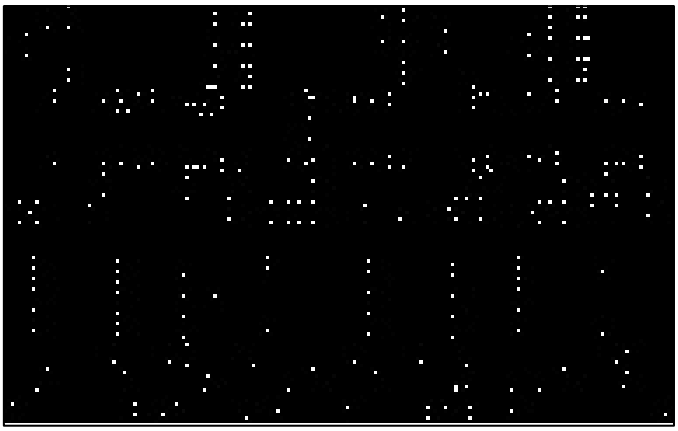

Fig. 3. Sample Output

The output of our Discriminator is a prediction map, wherein each unit represents a distinct patch within the se- quence. In other words, for every set of musical symbols with a length matching the patch size, there exists a corresponding value that predicts the authenticity of that patch. Specifically, this value indicates whether the patch adheres to the musical structure of the dataset being analysed. The utilisation of this methodology, frequently employed in the context of image-generating generative adversarial networks (GANs), guarantees the model's emphasis on local structural characteristics.

The present output function fulfils the objective of integrating prior knowledge pertaining to musical structure into our computational model. One prevalent approach to conceptualising the task of musical generation involves framing it as a language modelling problem. In this framework, each discrete symbol within the musical sequence is regarded as an individual word. These words are then combined to form phrases, periods, and other musical structures. In the context of this analogy, the objective of the proposed loss function is to determine the plausibility of individual sentences within a given text or, drawing a parallel to music, the realism of short musical motifs, such as phrases or partial phrases.

VI. EXPERIMENTATION

The first step involved generating latent points, where the latent dimension was set to 128, and these points were randomly generated Fig.3.

These latent points were then passed to the Generator, which used the power of BERT to generate a tensor array. To visualize the generated music, the tensor array was con- verted into an image representation. This image served as a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

Player	MOS
Human	4.5
BERT-GAN	3.5

Table I - Comparative Analysis

visual representation of the generated music, allowing for easy interpretation and analysis.

In order to further evaluate the music's quality and subjective characteristics, the image representation was converted into the Musical Instrument Digital Interface (MIDI) format.

The MIDI file was subsequently employed for the purpose of generating audio content, which was then subjected to a process of subjective evaluation. The utilisation of the mean opinion score (MOS) served as a quantitative measure for evaluating the perceived quality and musicality of the generated music.

VII. RESULTS

Various tests were done to thoroughly measure how well the music generation worked based on GANs and BERT. Initially, a comparison was made between the audio samples generated by the system and the ground truth. The ground truth is a reference point that represents the synthesized music and ideally aligns with the input guidelines. The comparison facilitated an evaluation of the precision of the system in replicating the intended musical output.

A Mean Opinion Score (MOS) [7] was computed to evaluate the quality of the produced music. A sample size of 50 participants with a non-musical background was selected at random to evaluate the authenticity of the audio using a rating scale ranging from 1 to 5. The assessment comprised a combination of human recordings, music produced through alternative techniques, and music produced by the GAN-BERT model under consideration. Each sample was equally represented. The presented results in Table. I suggest that the model has the ability to produce coherent music for specific input parameters. Additional improvements can be achieved by increasing the number of iterations used to train the model, resulting in an overall enhancement of its performance.

The GAN-BERT music generation system's performance was comprehensively understood through the experiments and evaluations conducted. The discoveries offer significant perspectives for the forthcoming enhancement and advancement of the suggested system.

REFERENCES

[1] R. R. Spangler, "Rule-based analysis and generation of music," en, Medium: PDF Version Number: Final, Ph.D. dissertation, California Institute of Technology, Mar. 2008. DOI: 10 . 7907 / YXTQ- 4057. [Online]. Available: https://resolver.caltech.edu/CaltechETD:etd-02262008- 114413 (visited on 12/27/2023).

[2] R. P. Whorley and D. Conklin, "Music Generation from Statistical Models of Harmony," en, Journal of New Music Research, vol. 45, no. 2, pp. 160–183, Apr. 2016, ISSN: 0929-8215, 1744-5027. DOI: 10.1080/09298215. 2016 . 1173708. [Online]. Available: https : / / www . tandfonline . com / doi / full / 10 . 1080 / 09298215 . 2016 . 1173708 (visited on 12/27/2023).

[3] I. J. Goodfellow, "Generative adversarial networks," arXiv.org, Jun. 2014. [Online]. Available: https : / / doi . org/10.48550/arXiv.1406.2661.

[4] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention Is All You Need," 2017, Publisher: arXiv Version Number: 7. DOI: 10 . 48550 / ARXIV . 1706 . 03762. [Online]. Available: https://arxiv.org/abs/1706.03762 (visited on12/27/2023).

[5] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang,"MuseGAN: Multi-track Sequential Generative Adver- sarial Networks for Symbolic Music Generation and Accompaniment," 2017, Publisher: arXiv Version Num- ber: 2. DOI: 10 . 48550 / ARXIV. 1709 . 06298. [Online]. Available: https://arxiv.org/abs/1709.06298 (visited on 12/27/2023).

[6] A. Dosovitskiy, "An image is worth 16x16 words: Trans- formers for image recognition at scale," arXiv.org, Oct. 2020. [Online]. Available: https : / / doi . org / 10 . 48550 /arXiv.2010.11929.

[7] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinionscore (MOS) revisited: Methods and applications, limita-tions and alternatives," en, Multimedia Systems, vol. 22, no. 2, pp. 213–227, Mar. 2016, ISSN: 0942-4962, 1432- 1882. DOI: 10 . 1007 / s00530 - 014 - 0446 - 1. [Online].Available: http : / / link . springer. com / 10 . 1007 / s00530 -014-0446-1 (visited on 12/27/2023).