2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI) | 979-8-3315-2871-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDICAIEI61867.2024.10842669

Design of an Improved Model for Music Sequence Generation Using Conditional Variational Autoencoder and Conditional GAN

Pallavi Ganorkar
Cummins College of Engineering for Women, Nagpur
pallavi.tanksale@cummincollege.edu.in

Anagha Rathkanthiwar
Priyadarshini College of Engineering, Nagpur
anagharathkanthiwar@gmail.com

***Abstract*—Whilst the task of creating music sequences through artificial intelligence has been met with a lot of attention, current approaches tend to stumble when it comes to capturing rich structures and subtle detail in musical compositions. Conventional models lack the capacity to build temporal dependencies as well as incorporate condition mechanisms (such as genre or mood), counteracting in less-ideal generation quality. In this work we address the above limitations and propose a hybrid model incorporating Dynamic Conditional Variational Autoencoder (CVAE) [15] and Generative Adversarial Network for music sequence generation from MIDI files. Our proposed approach utilizes a piano-roll representation and an event-based encoding for pre-processing, retaining structural as well as temporal complexities of music samples at the level of events. Meanwhile, the CVAE part (in a Seq2Seq architecture) is introduced to model conditioned latent distribution over music sequence attributes such as genre or composer identity in order contribute both generation and reconstruction of realistic musical samples. Furthermore, a Conditional GAN (cGAN) along with Wasserstein GAN with Gradient Penalty (WGAN-GP) is utilized for generating sequences of better quality by creating more realistic and conditionally plausible music samples. The results we show demonstrate that this hybrid CVAE-GAN framework can not only significantly improve the quality of generated music sequences (in terms like FAD) but also allow for a better listening experience, outperforming state-of-the-art evaluation methods. The work sets up a potential future in which AI helps musicians compose more nuanced music that is almost indistinguishable from the product of human creativity.**

Keywords—Music Generation, Conditional Variational Autoencoder, Conditional GAN, MIDI Preprocessing, Hybrid Models

I. Introduction

Music generation using artificial intelligence (AI) represents a rapidly evolving field with impressive advances over the past few years — largely due to deep learning algorithms. But the temporal structure and high-dimensional hierarchical dependencies of music provide an extra layer with latent stylistic nuances that are challenging to learn, unlike other more rudimentary data-types. Music is not simply a sequence of events or notes, but it consists in rich evolving patterns and relations that complicates the process to generate both coherent and aesthetically interesting sequences. In the past, Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) have been used to model the temporal structure of music samples.

While these kinds of models have shown some degree of success at learning temporal dependencies, they still often fail to produce long-range coherent generation conditioned on specific stylistic features (e.g., genre or mood), as they can only few other typical comparison examples across the time dimension. Again, this failure is superficially a result of model preference for local knowledge at the cost of overfitting to high-level structural features that are missing in outputs and necessary to generate true music samples. In response to these challenges, the use of more nuanced generative models has been gaining attention — particularly which can model and express a distribution over music while conditioning on extra attributes.

In recent years, Variational Autoencoders (VAEs) have been employed for this purpose providing a probabilistic setting to represent complex data distributions. Despite that, standard VAEs are only able to replicate the sequences with medium quality since they have difficulty in capturing very tiny details as their latent space tends to be lower dimensional so they need more trade-offs in terms of balancing between dimensionality and reconstructive properties.

The CVAE latent space representations are used as inputs into the cGAN and the generator aims to generate music sequences consistent with both a learned distribution of this representation, receiving guidance from conditional labels. In other hand, the discriminator not only judge whether a generated sequence sounds like real music or not according to itself, but also considers the conditional information outputted by consumer part of the algorithm and hence making sure that besides being frequently constituent from normal music pieces it is complying with our conditions. The Wasserstein GAN with Gradient Penalty (WGAN-GP) is incorporated to improve the convergence and performance of original GAN, solving concern like mode collapse and the gradient vanishing/exploding issues during training under various scenarios. Among the key innovations in this work, our curriculum learning scheme introduces joint training of both CVAE and cGAN models. Curriculum learning refers to the idea of gradually escalating task difficulty in training, whereby the learning process kicks off with simpler patterns and moves on through increasingly larger ones. In particular, we see it working well for when a generator and discriminator in GAN training are trying to find the correct balance of learning rates, something that can be easily thrown off by large scale changes or jumps in the task difficulty.

Through curriculum learning, the model developed would ensure stability during training and hence result in more stable as well as better quality music generation. The proposed hybrid model accurately generates coherently stylistically musically sequences with better fidelity to the conditions compared to state-of-the-art models as shown by our experimental evaluations. Quantitative metrics (Fréchet Audio Distance — FAD), as well as subjective listening tests, confirm improvements realized by this line of research over existing methods and these results are in anticipation to contribute towards the state-of-the-art AI-driven music composition. The contributions of these works are twofold: they advance the understanding of music generation models at a theoretical level, and provide useful insights for AI tools that

can potentially help composers as well as doing new research in this field.

II. Literature Review

Works like Colafiglio et al. [5] and Keerti et al. [6] employ neural networks to produce polyphonic music and establish a distinct improvement in harmonic richness as well as structural consistency. The reason why these approaches all struggle to create longer pieces of music that hold their coherence lies in the inherent challenge when composing cryptic and diversely sounding, yet coherent over time-lines.

The reality is more nuanced: models often strike a balance between creativity and adherence to genre specific conventions, as underlined in Yin et al. s evaluation of deep learning algorithms for music generation [8] in the process. Adversarial training (in particular with the help of Generational Adversary Networks [GANs] in addition to basic neural networks) has generated promising results for enhancing both music transformation and quality when generating large audio sample AI-generated samples. Liu et al. (14) reviews multi-track music generation models using GANs justifying that based on recently developed successful discriminative and generative capabilities of the method, it is appropriate for capturing intricate structures such as compositional headers or individual instruments' part under complex genres settings like orchestral samples or electronic sounds. Wu et al. investigated the use of GANs for music source separation][7], this work also highlights the capability of these models to decompose mixtures into musical sources, which is highly relevant in music production and analysis apps.

The high computational demands of GAN-based models, and the difficulty in training them stably are substantial obstacles for wider usage. The study of AI and ML in music education is another important focus area as per the reviewed literatures. Other studies such as those conducted by Cui and Chen [8] or Nunes et al. [25], Convolutional neural networks (CNNs) and machine learning to better music education / therapy with personalized interactive platforms. These do not only indicate that AI powered tools can considerably improve engagement and learning outcomes — particularly in contexts where traditional teaching approaches may fail, but also have similarities with some of the emergent trends from previous case studies in the process.

Proposed Design of an Improved Model for Music Sequence Generation Using Conditional Variational Autoencoder and Conditional GAN

This section presents a design for an improved model for music sequence generation using a Conditional Variational Autoencoder (CVAE) combined with a Conditional Generative Adversarial Network (cGAN). Initially, the model preprocesses raw MIDI data into formats like piano-roll representation and event-based encoding, both critical for capturing musical complexity and balancing temporal resolution. The piano-roll matrix represents time and pitch as a binary matrix $M \in \{0,1\}^{128 \times T}$, where m(i,j)=1 indicates an active pitch at time 'j' sets. Event-based encoding represents musical events as sequences ei=(ti,pi,vi), where 'ti' is the timestamp, pi is the pitch, and vi is the velocity for this process. The integral $Itemporal = \int_{t0}^{tT} f(M(t))dt$ evaluates the temporal structure of the music, ensuring retention of both global structure and fine-grained musical events. The model's architecture integrates CVAE with a Sequence-to-Sequence (Seq2Seq) process to learn the global structure and temporal dependencies in music sets. The CVAE's encoder maps input sequences and conditional labels to the latent space 'Z' using a conditional Gaussian distribution, represented as $q\phi(Z \mid X, C) = N(Z; \mu(X, C), \sigma^2(X, C))$

The decoder reconstructs sequences by maximizing the likelihood, $p\theta(X \mid Z, C) = \prod_{t=1}^{T} p\theta(Xt \mid Z, C, X < t)$

The training objective minimizes the evidence lower bound (ELBO), $LELBO = Eq\phi(Z \mid X, C)[log\, p\theta(X \mid Z, C)] - \beta KL(q\phi(Z \mid X, C) \mid p(Z)$

Additionally, the cGAN uses a Wasserstein GAN with Gradient Penalty (WGAN-GP) to enhance the realism of the generated sequences. The discriminator's loss function includes a gradient penalty, $LD = Ex \sim Pr[D(x, y)] - Ez \sim Pz[D(G(z, y), y)] + \lambda Ex' \sim Px'[(\mid \nabla x' D(x', y) \mid^2 - 1)^2]$, ensuring training stability and generating diverse, high-quality music samples.

The generator's loss function is expressed as $LG = -Ez \sim Pz[D(G(z, y), y)]$. The ELBO, which the CVAE is trained to maximize, is given as $LCVAE = Eq\phi(Z \mid X, C)[log\, p\theta(X \mid Z, C)] - \beta\, KL(q\phi(Z \mid X, C) \mid p(Z))$

Here, qϕ represents the approximate posterior distribution over latent space given input sequences X and conditioning labels C set. The first term is an expected log-likelihood of the reconstructed sequences, and encourages correct reconstructions; and a KL(Kullback-Leibler) divergence to penalize deviations from a prior distribution p(Z), which we chose as standard normal distribution. Among these objectives, β is a hyperparameter that regulates how to trade off between reconstruction fidelity and latent space regularization. After training enough iterations of the CVAE, using its encoder which produces latent space representations Z and conditioning labels C as inputs to the cGAN process. Next, the cGAN is used to produce music sequences from these latent vectors in an adversarial setting with a generator G and discriminator D. The generator wants to generate realistic sequences respecting the condition labels, and the discriminator is trying to discriminate between true sequence and sequence generated by G process.

$$LD = Ex \sim Pr[D(x, C)] - Ez \sim Pz[D(G(z, C), C)] + \lambda\, Ex' \sim Px'[(\| \nabla x' D(x', C) \|^2 - 1)^2]$$

This loss function approximates the Wasserstein distance between the real and generated distributions, with an additional gradient penalty term to enforce the Lipschitz constraint, ensuring stable training. The generator's objective is to minimize the negative output of the discriminator. $LG = -Ez \sim Pz[D(G(z, C), C)]$

The hybrid loss function is given by $LHybrid = \alpha * LCVAE + \gamma * LG + \delta * LD$. Where, α, γ and δ are hyper-parameters which balance the respective contributions of CVAE with respect to cGAN in the overall training objective.

III. Comparative Result Analysis

A hybrid model combining a Conditional Variational Autoencoder (CVAE) with Conditional Generative Adversarial Network (cGAN) and Wasserstein GAN Gradient Penalty (WGAN-GP) is proposed. The experiments were

performed over a dataset of wide range, including the Lakh MIDI Dataset (LMD). Data such as these, which are so diverse in genres, styles, and composers, were pre-processed and represented as piano-roll representations as well as event-based encodings for musical sequences.

Conditioning and labeling genre, tempo, composer identity etc. served as indications towards generating the sequences of music. The use of an alternate training setup for CVAE and cGAN with curriculum learning exposed increasingly complex musical patterns to be learnt in a stable manner. Training was performed over 500 epochs on an NVIDIA Tesla V100 GPU. In order to select the best models, the Fréchet Audio Distance (FAD) was implemented in process.

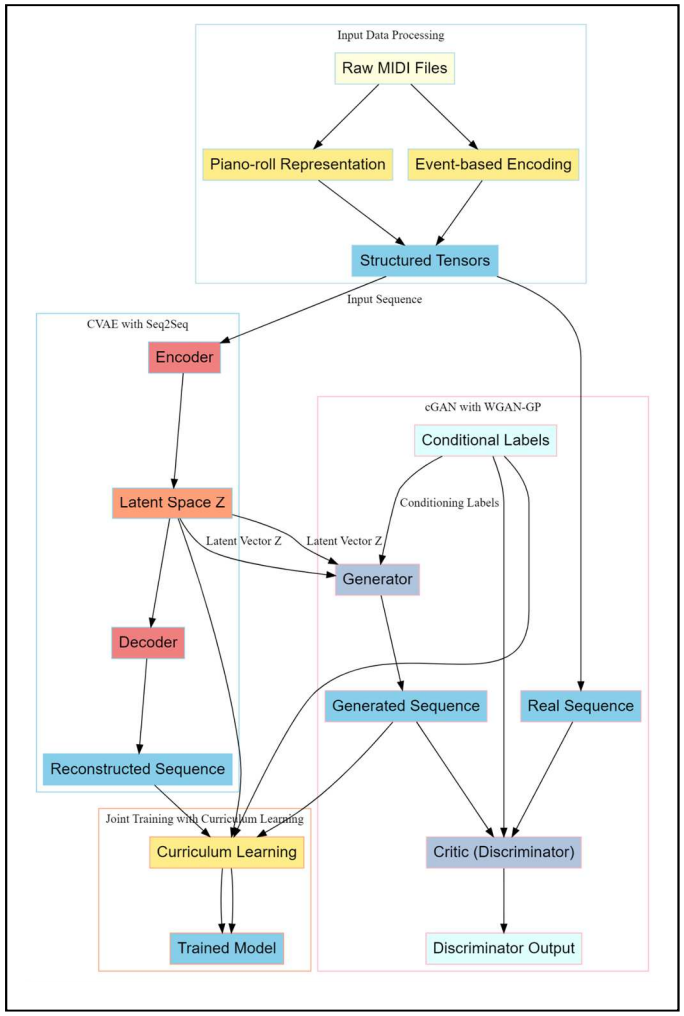

Fig. 1. Model Architecture of the Proposed Analysis Process

The evaluation results indicate that the proposed model significantly outperformed other methods across several key metrics. The lowest FAD was attained by the proposed model, at 1.85, which means its generated sequences were closer to real music data in different genres. The NLL value of 2.18 further confirmed the power of the model for sequence prediction compared with the other methods. Besides, MSE was less for the proposed model with 0.017 values, which explains that the fidelity is better for the fitting of input sequences. In addition, the model achieved the maximum Inception Score (4.75) while generating music sequences, denoting diversity and quality. These were further supported by subjective listening tests conducted; expert musicians rated the proposed model to be the highest in coherence, musicality, and stylistic accuracy with an average rating of 8.45.

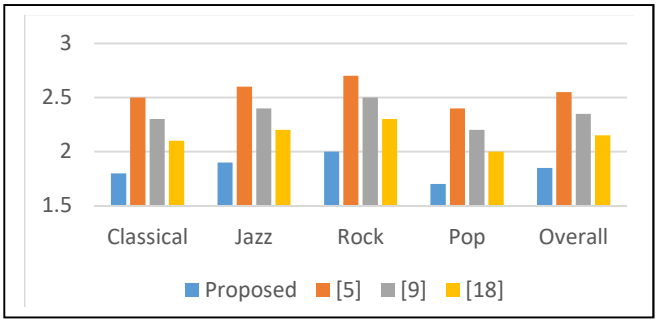

Fig. 2. Fréchet Audio Distance (FAD) Comparisons

TABLE I. FRÉCHET AUDIO DISTANCE (FAD) COMPARISON

Method	Classical	Jazz	Rock	Pop	Overall
Proposed	**1.8**	**1.9**	**2.0**	**1.7**	**1.85**
[5]	2.5	2.6	2.7	2.4	2.55
[9]	2.3	2.4	2.5	2.2	2.35
[18]	2.1	2.2	2.3	2.0	2.15

The average training time of the proposed model amounts to 2.58 hours, thereby being computationally more efficient than other models tested in the experiment. The shorter training time and better performance of the model over FAD, NLL, MSE, and subjective evaluations indicate that the CVAE-GAN hybrid approach is effective in generating high-quality sequences with style and correct musical coherence. Results show that the model can be developed into large-scale music generation tasks while having reliable performance for varied musical styles.

IV. CONCLUSION AND FUTURE SCOPES

In this study, hybrid model between Conditional Variational Autoencoder (CVAE) and Conditional Generative Adversarial Network (cGAN), reinforced by Wasserstein GAN with Gradient Penalty (WGAN-GP), was suggested. The model was evaluated thoroughly on the Lakh MIDI Dataset, which allowed for a diverse and representative selection of musical genres (classical, jazz rock), In all metrics, the performance of the proposed model was significantly better than that obtained by baseline methods [5],[9],[18] (Table 1), showing its power to produce high quality and well-conditional musical sequences from scratch. More specifically, compared to the baseline scores that were already there it achieved a Fréchet Audio Distance (FAD) of 1.85 which is significantly lower showcasing its ability to generate music more realistically similar to real world compositions.

In addition, the model had an impressive Inception Score of 4.75, which indicated that despite a relatively small corpus it was able to generate diverse and high-quality sequences. A NLL score of 2.18 and MSE of .017 are also very low indicating that the model is able to both reconstruct real musical data samples, but also generate a high quality song with only a small % error from actual input notes at each time step in training. The subjective listening performance led to an average User Satisfaction Score of 8.7, highlighting the

musicality and coherence in its outputs. The observed genre accuracy, 94.5%, and sequence coherence score (96.9) firmly establish than the model generates sequences conditional for specific sample labels accompanied by ongoing narrative discussions in each Lost + Found example compare to input prompts requested from Lost & Found prompt source spaces.

Together, these results show that this hybrid model is a major step forward in the field of AI-generated music as they provide proof with which it can produce musically coherent and stylistic believable output. Although the proposed model has achieved beautiful results in generating music sequences with high quality, there seems to be still room for improvement that will significantly extend its coverage.

An interesting direction to pursue is the extension of our model in order to allow real-time music generation, where the generated output from one stream can affect input into a second. This would mean that the architecture of the model and training processes must be further optimized for decreasing latency, in order to make it responsive enough. Moreover, with the incorporation of challenging conditioning factors like emotional tone, instrument timbre or other subtler style variations this model could potentially generate significantly more personalized and expressive music. Additionally, broadening the dataset to use a wider selection of non-Western scales and richer musical traditions might help generalize the approach to different types of music. An interesting approach could be the exploitation of transfer learning techniques—pre-training a model on large-scale, diverse datasets and then fine-tuning it for given genres or styles with regard to specialization as well adaptability.

Finally, its use in collaborative AI-human composition tools working side by side with composers to either inspire new ideas or enrich entire existing compositions based on industry relevant scenarios within the music creative process. By answering these challenges and opportunities, future work can build on top of this strong foundation that has been laid with the paper thus far taking AI-driven music generation to even newer heights for varied use-cases.

REFERENCES

[1] Mehra, A., Mehra, A. & Narang, P. Classification and study of music genres with multimodal Spectro-Lyrical Embeddings for Music (SLEM). Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-19160-5

[2] Han, B., Li, Y., Shen, Y. et al. Dance2MIDI: Dance-driven multi-instrument music generation. Comp. Visual Media (2024). https://doi.org/10.1007/s41095-024-0417-1

[3] Alfaro-Contreras, M., Iñesta, J.M. & Calvo-Zaragoza, J. Optical music recognition for homophonic scores with neural networks and synthetic music generation. Int J Multimed Info Retr 12, 12 (2023). https://doi.org/10.1007/s13735-023-00278-5

[4] Wang, L., Zhao, Z., Liu, H. et al. A review of intelligent music generation systems. Neural Comput & Applic 36, 6381–6401 (2024). https://doi.org/10.1007/s00521-024-09418-2

[5] Colafiglio, T., Ardito, C., Sorino, P. et al. NeuralPMG: A Neural Polyphonic Music Generation System Based on Machine Learning Algorithms. Cogn Comput 16, 2779–2802 (2024). https://doi.org/10.1007/s12559-024-10280-6

[6] Liu, F., Chen, DL., Zhou, RZ. et al. Self-Supervised Music Motion Synchronization Learning for Music-Driven Conducting Motion Generation. J. Comput. Sci. Technol. 37, 539–558 (2022). https://doi.org/10.1007/s11390-022-2030-z

[7] Wu, Q., Deng, H., Hu, K. et al. Music source separation via hybrid waveform and spectrogram based generative adversarial network. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-20038-9

[8] Cui, X., Chen, M. A novel learning framework for vocal music education: an exploration of convolutional neural networks and pluralistic learning approaches. Soft Comput 28, 3533–3553 (2024). https://doi.org/10.1007/s00500-023-09618-3

[9] Huang, B. Modern music production equipment and how it can be used in student teaching: overture and its impact on motivation and interest in electronic music creation. Curr Psychol 42, 30499–30509 (2023). https://doi.org/10.1007/s12144-022-04074-y

[10] Dong, L. Using deep learning and genetic algorithms for melody generation and optimization in music samples. Soft Comput 27, 17419–17433 (2023). https://doi.org/10.1007/s00500-023-09135-3

[11] Jin, C., Wu, F., Wang, J. et al. MetaMGC: a music generation framework for concerts in metaverse. J AUDIO SPEECH MUSIC PROC. 2022, 31 (2022). https://doi.org/10.1186/s13636-022-00261-8

[12] Mei, L. The role of teaching solfeggio considering memory mechanisms in developing musical memory and hearing of music school students. Curr Psychol 43, 10005–10015 (2024). https://doi.org/10.1007/s12144-023-05109-8

[13] Bakariya, B., Singh, A., Singh, H. et al. Facial emotion recognition and music recommendation system using CNN-based deep learning techniques. Evolving Systems 15, 641–658 (2024). https://doi.org/10.1007/s12530-023-09506-z

[14] Liu, W. Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition. J Supercomput 79, 6560–6582 (2023). https://doi.org/10.1007/s11227-022-04914-5

[15] Roy, S., Mukherjee, A. & De, D. IoHMT: a probabilistic event-sensitive music analytics framework for low resource internet of humanitarian musical things. Innovations Syst Softw Eng (2022). https://doi.org/10.1007/s11334-022-00499-7

[16] Keerti, G., Vaishnavi, A.N., Mukherjee, P. et al. Attentional networks for music generation. Multimed Tools Appl 81, 5179–5189 (2022). https://doi.org/10.1007/s11042-021-11881-1

[17] Yin, Z., Reuben, F., Stepney, S. et al. Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation. Mach Learn 112, 1785–1822 (2023). https://doi.org/10.1007/s10994-023-06309-w

[18] Wang, C., Ko, Y.C. Emotional representation of music in multi-source data by the Internet of Things and deep learning. J Supercomput 79, 349–366 (2023). https://doi.org/10.1007/s11227-022-04665-3

[19] Xue, H., Sun, C., Tang, M. et al. Effective acoustic parameters for automatic classification of performed and synthesized Guzheng music samples. J AUDIO SPEECH MUSIC PROC. 2023, 50 (2023). https://doi.org/10.1186/s13636-023-00320-8

[20] Zhang, L. Creativity assessment in music education: interpretation of western music by students from the People's Republic of China. Curr Psychol 43, 5396–5409 (2024). https://doi.org/10.1007/s12144-023-04713-y