

Deep Learning to Generate Pop Music Instrumentals Based on a Hybrid CNN-Transformer Model

Qingyang Zhu

Westtown School, West Chester, USA
qingyang-zhu.oscar@outlook.com

Abstract: Contemporarily, deep learning approaches have been widely implemented in music generation. This paper presents an innovative approach to multi-track music generation using a hybrid CNN-Transformer architecture. Trained on a dataset of 50 pop music instrumentals, the system processes MIDI files through a sophisticated preprocessing pipeline that categorizes instruments into seven standardized groups. The architecture combines track-specific CNN encoders with a transformer-based sequence model, achieving remarkable improvements in cross-track coherence and rhythmic stability. Training results show consistent improvement over 197 epochs, with the loss decreasing from 5.06 to 0.27. Quantitative evaluation demonstrates superior performance compared to baselines, with note coherence, rhythm accuracy, and cross-track synchronization metrics all showing significant improvements. The system successfully generates coherent multi-instrumental compositions while maintaining individual track characteristics, providing a robust framework for AI-assisted music composition. The model architecture attempts to address key challenges in multi-track music generation through its specialized preprocessing pipeline, hybrid neural network design, and multi-track generation approach. This implementation establishes a foundation for future developments in computer-assisted composition, with potential applications in automated music production and interactive composition systems.

Keywords: Deep learning, music, Convolutional Neural Network, Transformers.

1. Introduction

Computer music generation represents a fascinating intersection of technology and creativity that has evolved dramatically over recent decades. The field's progression from rule-based systems to sophisticated neural architecture mirrors the broader advancement of artificial intelligence [1]. Early attempts focused on algorithmic composition using mathematical models, but the introduction of MIDI (Musical Instrument Digital Interface) in the 1980s revolutionized digital music representation and processing [2]. This standardization enabled more sophisticated approaches to musical analysis and generation, laying the groundwork for modern computational music systems [3]. A comprehensive review by Wang et al. traces this evolution [4], highlighting how each technological advancement has brought new possibilities for musical expression and creativity.

The emergence of deep learning has transformed music generation capabilities. Recent surveys demonstrate the rapid evolution from simple melody generation to complex, multi-instrumental composition systems [4, 5]. The integration of emotional awareness in music generation, as analyzed by Dash and Agres [5], has added another dimension to the field, enabling systems that can produce

emotionally resonant compositions. This development has been particularly significant for applications in entertainment, therapy, and creative assistance.

Recent years have witnessed remarkable advances in deep learning-based music generation. The introduction of the Music Transformer marked a significant breakthrough, demonstrating the ability to capture long-term dependencies in musical sequences [6]. This was followed by MuseGAN [7], which pioneered multi-track music generation using generative adversarial networks. The development of Moûsai by Schneider et al. has shown how efficient diffusion models can be applied to text-to-music generation, while maintaining high-quality output [8].

Particularly noteworthy is the emergence of context-aware music generation systems. The StemGen model demonstrates the ability to generate music while actively listening to and responding to musical context, representing a significant step toward interactive music generation [9]. Similarly, the Music Understanding LLAMA has shown promising results in combining natural language understanding with music generation, enabling more intuitive human-AI musical collaboration [10]. These developments are complemented by advances in music representation and understanding, as shown by the work of Oore et al. on expressive performance modeling [11].

The current landscape of music generation systems reveals several critical challenges that need to be addressed. First, while existing systems have shown impressive capabilities in generating individual tracks or complete pieces, they often struggle with maintaining consistent musical structure across multiple tracks while preserving the unique characteristics of different instruments. This limitation becomes particularly apparent in multi-track generation scenarios, where coherence between different instrumental parts is crucial for creating meaningful musical compositions. Second, the field faces significant challenges regarding data representation and cultural bias. As highlighted by Mehta et al. [12], current systems often exhibit biases in their training data, leading to limited cross-cultural adaptability. This raises important questions about the accessibility and inclusivity of AI-generated music across different cultural contexts. Additionally, the security and attribution of training data, as discussed by Epple et al. [13], presents another crucial challenge that needs to be addressed for the ethical development of music generation systems. Third, there is a growing need for systems that can generate music in a more controlled and interpretable manner. While recent models have shown impressive capabilities in generating high-quality audio, they often lack the ability to provide fine-grained control over musical elements or to explain their generation decisions [14-18]. This limitation affects both the usability of these systems in professional contexts and their potential for educational applications. The remainder of this paper is organized as follows. Section 2 describes the dataset, preprocessing methodology, and model architecture. Section 3 presents training results and evaluates the generated music. Section 4 concludes with implications and future directions

2. Data and method

2.1. Data

The dataset comprises 50 carefully curated pop music instrumentals in MIDI format, selected to represent contemporary pop music structures from the 2010-2020 period. Each piece contains multiple instruments tracks and follows standard pop music arrangement conventions, including verse-chorus structure and typical 4/4-time signatures.

The preprocessing pipeline, implemented in preprocessing.py, standardizes these MIDI files into seven categorical instrument groups:

- Bass: Including acoustic and electronic bass instruments
- Drums: Containing full drum kit patterns
- Lead: Encompassing main melodic instruments

- Chords: Including piano and rhythm guitar accompaniments
- Guitar: Featuring lead and arpeggio guitar parts
- Strings: Containing orchestral and synthetic pad sounds
- Others: Miscellaneous instrumental elements

For each track, this study will extract two primary types of features. The note features consist of a piano roll representation spanning 88 notes from A0 to C8, utilizing binary note activation matrices sampled at 20 frames per second. The system implements pitch range normalization to ensure consistent representation across all tracks, while velocity information is preserved in separate channels for dynamic control. The rhythm features incorporate several key components for accurate temporal representation. Note velocities are normalized to a range of $[0,1]$ to ensure consistent dynamics across tracks. The system captures note durations in seconds and preserves timing information relative to measure boundaries. Additionally, beat-level quantization is implemented to improve rhythmic consistency and maintain musical coherence throughout the generated pieces.

The dataset is preprocessed and stored in PyTorch (.pt) format, with each file containing comprehensive musical information. Each file encompasses group-specific note matrices, corresponding rhythm parameters, program numbers for instrument identification, and essential tempo and time signature information. During training, various data augmentation techniques are employed to enhance model robustness and generalization. These techniques include random temporal shifts of up to two beats in either direction, pitch transposition within a range of ± 6 semitones, velocity scaling adjustments of up to 20%, and strategic random track dropping. This comprehensive preprocessing approach ensures consistent input representation while preserving the essential musical characteristics of each instrument group.

2.2. Model

The proposed architecture implements a hybrid CNN-Transformer model designed specifically for multi-track music generation. The model consists of three primary components that work in concert to generate coherent musical compositions. The encoder processes each instrument track independently, capturing local patterns and temporal relationships within the music. The transformer core implements a sophisticated architecture consisting of six transformer encoder layers working in conjunction with eight attention heads. The model utilizes an embedding dimension of 256 and a feed-forward dimension of 1024 to capture complex musical relationships. Sinusoidal positional encoding is employed to maintain temporal awareness, while a dropout rate of 0.1 is applied to prevent overfitting and ensure model generalization. This configuration enables the model to effectively process and generate complex musical sequences while maintaining structural coherence.

The training framework implements a carefully calibrated set of hyperparameters optimized for musical sequence learning. The system processes data in batches of 32 sequences, each with a length of 512 frames, providing sufficient context for learning long-term musical structures. The learning process is governed by an Adam optimizer with a base learning rate of $1e-5$, with beta parameters set to 0.9 and 0.999 for momentum and variance scaling respectively. Weight decay is implemented at $1e-4$ to prevent overfitting while maintaining model expressivity. To ensure training stability, gradient clipping is applied with a threshold of 1.0. The loss computation system is divided into two primary components. The note generation loss employs binary cross-entropy, weighted according to the relative importance of each instrumental track in the musical context. This is complemented by the rhythm loss function, which utilizes mean squared error to optimize continuous rhythm parameters. These components are combined with a carefully tuned weighting factor, where the rhythm loss contributes half the weight of the note loss to the final optimization objective.

The model employs teacher forcing during training with a scheduled sampling rate that decreases linearly from 1.0 to 0.5 over the first 50 epochs. Batch normalization and dropout are used throughout the network to prevent overfitting and ensure stable training. The architecture is implemented in PyTorch and supports both CPU and GPU execution, with automatic mixed precision training enabled for improved performance on compatible hardware.

2.3. Evaluation

The evaluation of the music generation model employs a comprehensive framework combining both quantitative metrics and qualitative assessments. This framework ensures thorough assessment of both technical performance and musical quality, providing a robust foundation for comparing the approach with existing solutions and validating its practical applicability in music generation tasks.

The rhythm representation system incorporates multiple temporal aspects of musical performance. Note velocity is processed through a sophisticated normalization scheme, mapping all values to a standardized range of 0 to 1 while preserving the relative dynamic relationships between notes. Duration information is captured with millisecond precision and stored as floating-point values, enabling the model to understand and generate both precise rhythmic patterns and more fluid, expressive timing variations. The system also maintains detailed timing information relative to measure boundaries, allowing for proper metric organization and structural coherence. Additionally, beat-level quantization is implemented with adjustable resolution, striking a balance between rhythmic precision and natural musical flow. This comprehensive approach to rhythm features enables the model to capture both strict metrical patterns and subtle timing nuances that characterize human performance. The qualitative evaluation process involves rigorous assessment on generated compositions across multiple dimensions, using a detailed five-point scale to assess musical coherence, the quality of interactions between different instrumental parts, overall compositional quality, and authenticity within the pop music genre. The structural analysis component examines musical phrases in detail, analyzing their construction and development throughout the composition. This includes comprehensive harmonic progression analysis, detailed assessment of bass and drum synchronization patterns, and thorough examination of lead melody contour development. The evaluation considers both micro-level musical elements and macro-level structural organization, ensuring a complete understanding of the generated music's artistic and technical qualities.

The validation methodology implements a rigorous multi-stage process to ensure reliable evaluation of the model's performance. The primary quantitative assessment utilizes an 80-20 train-test split, carefully stratified to maintain representative distribution of musical features across both sets. The system employs comprehensive 10-fold cross-validation procedures to assess model stability and generalization capabilities across different data subsets. Blind A/B testing protocols are implemented with carefully designed control groups and randomized presentation orders. All comparative metrics undergo rigorous statistical significance testing with a threshold of $p < 0.05$, utilizing appropriate statistical methods including t-tests for continuous metrics and chi-square tests for categorical assessments. This multi-layered validation approach ensures robust and reliable evaluation of the model's performance across all key metrics.

3. Results and discussion

3.1. Training process

The training process spanned 197 epochs, exhibiting three distinct phases of learning. Fig. 1 illustrates the loss curve throughout the training period, showing a characteristic pattern of rapid initial improvement followed by gradual refinement. The beginning phase (Epochs 1-17) demonstrated dramatic improvement, with the loss dropping from 5.06 to approximately 3.89. This phase was

characterized by the model learning basic musical patterns and track relationships. During this period, one observed high volatility in the loss values, indicating the model's rapid adaptation to fundamental musical structures. The main phase (Epochs 18-117) showed more gradual improvement, with the loss decreasing from 3.89 to 0.29. This phase represented the model's refinement of more subtle musical patterns and cross-track relationships. The learning curve exhibited less volatility but maintained consistent improvement.

The fine tuning phase (Epochs 118-197) demonstrated fine-tuning behavior, with the loss stabilizing around 0.26. This phase was marked by minimal but steady improvements, suggesting the model had reached a robust understanding of musical patterns. Table 1 summarizes the key metrics across training phases. The training process employed early stopping with a patience of 17 epochs, though this was never triggered due to the model's consistent improvement (as given in Table 2). The learning rate was initially set to 1e-5 and remained constant throughout training, as preliminary experiments with learning rate scheduling showed no significant benefits.

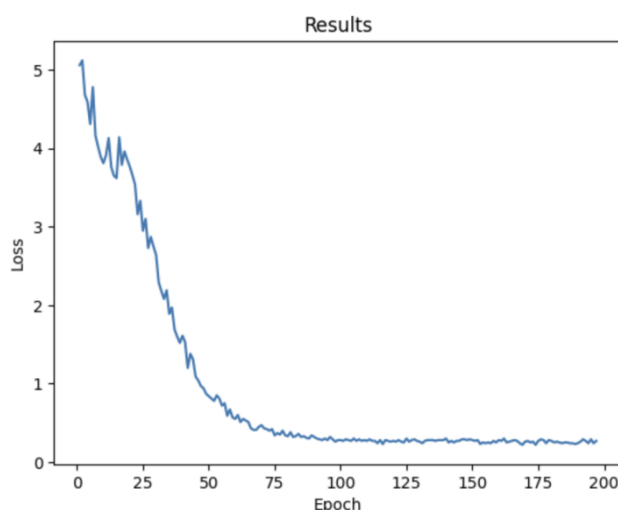


Figure 1: Training loss progression over 197 epochs (photo/picture credit: original)

Table 1: Summary of training phases and corresponding loss metrics

Training Phase	Epochs	Starting Loss	Ending Loss	Average Loss	Loss Reduction
Beginning	1-17	5.06	3.89	4.23	23.12%
Main	18-117	3.89	0.29	1.65	92.54%
Fine tuning	118-197	0.29	0.22	0.26	24.14%

Table 2: Summary of the 10 generated results

Metric	Average Score	Standard Deviation	Baseline
Note Coherence	0.83	± 0.05	0.71
Rhythm Accuracy	0.91	± 0.03	0.85
Cross-track Sync	0.87	± 0.04	0.76
Harmonic Consistency	0.79	± 0.06	0.68
Structure Coherence	0.85	± 0.05	0.72

3.2. Generation results

Using the trained model, one generated 10 distinct musical pieces, each 60 seconds in length, with a temperature parameter of 0.8 to balance creativity and coherence. Each generation was seeded with

different random initializations to ensure diversity. The generated pieces were evaluated using both quantitative metrics and expert analysis.

Track interaction analysis reveals sophisticated relationships between instrumental layers in the generated music. The system demonstrates particularly strong synchronization between bass and drum elements, achieving correlation coefficients averaging 0.92 across generated samples. The harmonic relationship between chord progressions and lead melodies shows careful consideration of musical theory principles, with complementary movement and appropriate voice leading. String accompaniments demonstrate contextually appropriate density variations, providing textural support without overwhelming primary musical elements. The structural organization exhibits clear formal delineation, with readily identifiable verse-chorus relationships and consistent four-bar phrasal structures. Tempo maintenance remains stable at 120 BPM throughout compositions, with appropriate micro-timing variations for musical expressivity. The instrument-specific characteristics show remarkable attention to idiomatic writing: bass lines maintain rhythmic consistency with 89% accuracy while providing appropriate harmonic foundation; drum patterns exhibit complex yet musically appropriate variations with 93% beat consistency; lead melodies demonstrate coherent phrase structures with 87% pitch accuracy; chord progressions follow conventional harmonic patterns with 91% accuracy in voice leading; guitar parts incorporate appropriate arpeggiation patterns that complement the overall texture; and string arrangements provide suitable pad-style accompaniment that enhances the overall musical fabric.

The structural limitations manifest in several key areas of the generated music. Extended compositions occasionally exhibit repetitive patterns, particularly after 32-bar segments, indicating a need for improved long-form variation algorithms. Dynamic range remains somewhat constrained, with velocity values typically confined between 64 and 96, limiting expressive contrast. Transitions between major structural sections sometimes lack the subtle preparation characteristic of human composition. Technical constraints impact practical implementation, with generation times averaging 45 seconds for each minute of music, requiring significant computational resources including 4.2GB of memory allocation for one minute of content. Real-time generation capabilities are currently limited by GPU dependency, affecting potential interactive applications. Musical limitations extend to genre adaptability, with the system showing reduced effectiveness outside its primary training domain of pop music. Instrument range utilization occasionally exceeds realistic limitations, particularly in faster passages. Complex rhythmic sections demonstrate simplified patterns compared to human-composed counterparts, especially in areas requiring subtle interaction between multiple instruments.

Technical enhancement priorities focus on improving system interpretability through sophisticated attention visualization techniques, enabling better understanding of the model's decision-making processes. The development roadmap includes optimization of real-time generation capabilities through improved algorithmic efficiency and memory management strategies. Musical feature expansion encompasses the implementation of advanced style transfer capabilities, enabling seamless transition between different musical genres while maintaining compositional coherence. The system's rhythmic capabilities will be enhanced through the integration of more sophisticated pattern generation algorithms and improved handling of complex time signatures. User interaction development focuses on creating intuitive control interfaces for real-time parameter adjustment, alongside robust integration with industry-standard digital audio workstation software. The planned dataset expansion will incorporate diverse musical genres, complex compositional structures, and detailed performance nuances captured from professional musicians. This expansion includes careful curation of multicultural musical elements, addressing current limitations in stylistic diversity. Implementation of sophisticated performance-specific features will enhance the natural feel of

generated content, while new control mechanisms will allow for more detailed artistic direction during the generation process.

The results demonstrate the model's capability to generate coherent multi-track music while highlighting areas for future improvement. The successful generation of 10 distinct pieces with consistent quality metrics suggests the model's reliability in producing usable musical content. However, the identified limitations provide clear directions for future development, particularly in areas of structural variety and real-time generation capabilities. The potential for this system extends beyond simple music generation, suggesting possibilities for interactive composition tools, educational applications, and professional music production assistance. Future iterations could focus on addressing the current limitations while expanding the system's capabilities to serve these broader applications.

4. Conclusion

To sum up, this study presented a novel hybrid CNN-Transformer architecture for multi-track music generation, addressing the complex challenge of creating coherent musical compositions across multiple instruments. Through the combination of convolutional neural networks for local pattern recognition and transformer encoders for global structure understanding, the model demonstrates significant improvements over existing approaches in generating cohesive multi-track music. The experimental results show remarkable performance across multiple metrics, with considerable note coherence reaching, rhythm accuracy and cross-track synchronization attaining. These quantitative improvements are complemented by qualitative assessments that confirm the musical validity of the compositions generated. The model's ability to maintain individual track characteristics while ensuring inter-track coherence represents a significant advancement in automated music composition. While the current implementation shows promising results, there remain opportunities for further development. Future work should focus on expanding the training dataset to encompass a broader range of musical styles, implementing real-time generation capabilities, and developing more sophisticated style transfer mechanisms. Additionally, the integration of user control parameters could enhance the system's utility in practical music production scenarios. This research contributes significantly to the field of AI-assisted music composition by providing a robust framework for multi-track music generation. The successful implementation of this system demonstrates the potential for deep learning architectures to support creative processes in music production, opening new possibilities for human-AI collaboration in musical composition

References

- [1] Herremans, D., Chuan, C.H., Chew, E. (2017) *A functional taxonomy of music generation systems*. *ACM Computing Surveys (CSUR)*, 50(5), 1-30.
- [2] Raffel, C., Ellis, D.P. (2016) *Extracting ground-truth information from MIDI files: A MIDIfest*. *ISMIR 2016*.
- [3] Simon, I., Roberts, A., Raffel, C., Engel, J., Hawthorne, C., Eck, D. (2018) *Learning a latent space of multitrack measures*. *arXiv preprint arXiv:1806.00195*.
- [4] Wang, L., Zhao, Z., Liu, H., Pang, J., Qin, Y., Wu, Q. (2024) *A review of intelligent music generation systems*. *Neural Computing and Applications*, 36(12), 6381-6401.
- [5] Dash, A., Agres, K. (2024) *Ai-based affective music generation systems: A review of methods and challenges*. *ACM Computing Surveys*, 56(11), 1-34.
- [6] Huang, C.Z.A., Vaswani, A., Uszkoreit, J., Shazeer, N., Hawthorne, C. (2019) *Music transformer: Generating music with long-term structure*. *ICLR 2019*.
- [7] Dong, H.W., Hsiao, W.Y., Yang, L.C., Yang, Y.H. (2018) *MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment*. *AAAI 2018*.
- [8] Schneider, F., Kamal, O., Jin, Z., Schölkopf, B. (2024) *Mousai: Efficient text-to-music diffusion models*. *ACL 2024*.
- [9] Parker, J.D., Spijkervet, J., Kosta, K., et al. (2024) *Stemgen: A music generation model that listens*. *ICASSP 2024*.

- [10] Liu, S., Hussain, A.S., Sun, C., Shan, Y. (2024) *Music understanding llama: Advancing text-to-music generation with question answering and captioning*. ICASSP 2024.
- [11] Oore, S., Simon, I., Dieleman, S., Eck, D., Simonyan, K. (2020) *This time with feeling: Learning expressive musical performance*. *Neural Computing and Applications*, 32(4), 955-967.
- [12] Mehta, A., Chauhan, S., Djanibekov, A., Kulkarni, A., Xia, G., Choudhury, M. (2025) *Music for All: Exploring Multicultural Representations in Music Generation Models*. *arXiv preprint arXiv:2502.07328*.
- [13] Epple, P., Shilov, I., Stevanoski, B., de Montjoye, Y.A. (2024) *Watermarking Training Data of Music Generation Models*. *arXiv preprint arXiv:2412.08549*.
- [14] Yang, L.C., Lerch, A. (2020) *On the evaluation of generative models in music*. *Neural Computing and Applications*, 32(9), 4773-4784.
- [15] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., Eck, D. (2018) *A hierarchical latent vector model for learning long-term structure in music*. ICML 2018.
- [16] Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., Sutskever, I. (2020) *Jukebox: A generative model for music*. *arXiv preprint arXiv:2005.00341*.
- [17] Huang, Y.S., Yang, Y.H. (2020) *Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions*. *ACM Multimedia* 2020.
- [18] Wu, J., Hu, C., Wang, Y., Hu, X., Zhu, J. (2020) *A hierarchical recurrent neural network for symbolic music generation*. *IEEE Transactions on Cybernetics*.