Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

AI-Driven ChatGPT Applications for Enhancing Music Education

Xiaofan Sun*

Department of Computer Science and Technology, Shandong University of Science and Technology, Qingdao, China

*Corresponding author: jiyi@ldy.edu.rs

Abstract. Dut to the limitations inherent in traditional music education approaches, the remarkable predictive and analytical prowess of artificial intelligence has been integrated to amplify the quality of music education. This paper investigates the integration of ChatGPT, an emerging machine learning technology based on deep learning and Natural Language Processing, with music education, and analyzes various approaches. These include the utilization of contemporary spectral analysis techniques to deconstruct songs and the adoption of three deep neural network models: SIConvNe, JointEmbedNe, and DistMatNet, to evaluate music. By combining the GPT architecture with Constraint Satisfaction Programming, ABC notation, and Domain-Specific Language technology frameworks, differentiated music exercises are generated. By employing a Large Language Model, centered around the LLaMA2 model, and refining its capabilities through ChatGPT pre-training and fine-tuning, the understanding and generation of music are continually enhanced. Finally, to address the challenges of inadequate explainability, algorithmic limitations, and low applicability in this field, this study discussed the combination of domain expertise with expert systems for model training, and employs Layer-wise Relevance Propagation and Local Interpretable Model-Agnostic Explanations to increase system transparency. Integrating audio analysis and visualization technologies, expanding system capabilities to accommodate a broader range of exercises, and developing algorithms to facilitate more interactive and personalized learning tools all serve to bolster the system's adaptability and flexibility. Utilizing domain-adaptive techniques, transfer learning, and multi-task learning methods, the GPT model's adaptability and generalization capability are strengthened, enabling it to skillfully generate pieces in varying styles based on a broader music corpus.

Keywords: ChatGPT; Deep Neural Network; Large Language Model; Transfer learning.

1. Introduction

Music, as a unique art form, plays a significant role in people's daily lives. It possesses intrinsic value and profound impact, resonating with the audience's emotions, cleansing the mind, and purifying thoughts. Appreciating music can help people relax, alleviate stress, and even ignite creativity while providing new perspectives and broader imaginative spaces, gradually becoming an essential part of spiritual and cultural life. As society progresses and people's living standards improve, the position of music education in the modern education system continues to rise. However, traditional music education methods have certain limitations to some extent. Issues such as insufficient personalized learning, low teaching efficiency, uneven teacher quality, and high education costs make it difficult to meet the ever-increasing educational demands. Additionally, for music educators, generating unbiased and effective song assessments is challenging and time-consuming. In this case, the application of artificial intelligence can be considered as a potential solutions due to its excellent prediction performance.

In recent years, deep learning has made significant progress in the field of artificial intelligence. In computer vision, significant advancements have been made in image recognition, object detection, and image generation, with Convolutional Neural Networks (CNNs) becoming the core technology in this domain. Meanwhile, Generative Adversarial Networks (GANs) have achieved remarkable results in image generation, style transfer, and denoising tasks. In speech recognition and synthesis, the application of Recurrent Neural Networks (RNNs) and Long Short-term Memory (LSTM)

networks has improved the quality and accuracy of speech recognition and generation tasks. Various deep learning models have provided key technologies for autonomous driving systems, such as environmental perception, decision-making and planning, and road object detection and recognition, significantly contributing to the development of automotive navigation and self-driving technologies. In the field of Natural Language Processing (NLP), breakthrough progress has been made as well. The rise of ChatGPT has sparked a new wave, as it is a chatbot based on the Generative Pre-trained Transformer (GPT) series of models. Building on NLP and deep learning technologies, ChatGPT provides users with intelligent, interactive language generation and understanding capabilities. ChatGPT can be applied to various scenarios, including intelligent customer service, online tutoring, image generation, and language learning. In the education sector, it has already made its mark in areas such as English teaching and programming education. However, music education has received little attention to date, which presents an opportunity for further exploration and potential applications in the future. Recently, Yuan et al. presented an open-source Large Language Model (LLM) with built-in musical capabilities, leveraging the LLaMA2 architecture for versatile music generation [1]. This approach enables personalized music scores for students while fostering their creativity. The resulting technology can notably enhance music education quality and efficiency. In a distinct study, Merino developed a deep-learning-based tool for melody dictation exercises, leveraging the GPT architecture and the Nottingham dataset [2]. This innovative solution has made a tangible impact on music education, offering convenience for learners and enhancing the field. In the research by Guo et al. [3], a music generation and evaluation method based on transfer learning was introduced, which converts musical melodies into textual representations and employs a pre-trained GPT-2 model for melody generation. Furthermore, integrating mathematical statistics and music theory, an objective music evaluation approach (MEM) was proposed for accurately assessing the quality of the generated melodies. In the research by Imasato et al. [4], the OpenAI GPT model was also employed as the foundation. To achieve control over the emotional aspect of generated music, they used a small discriminator model along with a method called "Plug and Play Language Models" (PPLM). Without requiring retraining of the large-scale model, this approach allowed the generation of music with specific emotional characteristics based on predefined emotional labels. Clearly, as demonstrated through various studies mentioned above, the development of AI technology has provided new opportunities for the integration of music education and ChatGPT, paving the way for complementary innovations in this field.

The remainder of this paper is organized as follows: First, in Section 2, this paper will present multiple approaches to promoting music education through ChatGPT. Next, in Section 3, this paper will delve into the current limitations of integrating music education with ChatGPT, as well as future directions for development. Finally, Section 4 will summarize the paper and offer a discussion of the conclusions drawn.

2. Method

2.1. Introduction of ChatGPT

The Transformer neural network architecture serves as the underpinning of ChatGPT (Fig. 1), a sophisticated language model proficient in comprehending and generating human-like responses. This architecture employs self-attention mechanisms, thereby empowering the model to effectively capture long-range dependencies and context within textual data. Through parallel processing in multiple encoder and decoder layers, the Transformer handles information with remarkable efficiency, rendering it well-suited for extensive language modeling tasks. Consequently, ChatGPT generates coherent, context-sensitive, and grammatically accurate responses, establishing itself as a formidable resource for an array of applications, including conversational agents, content generation, and problem-solving.

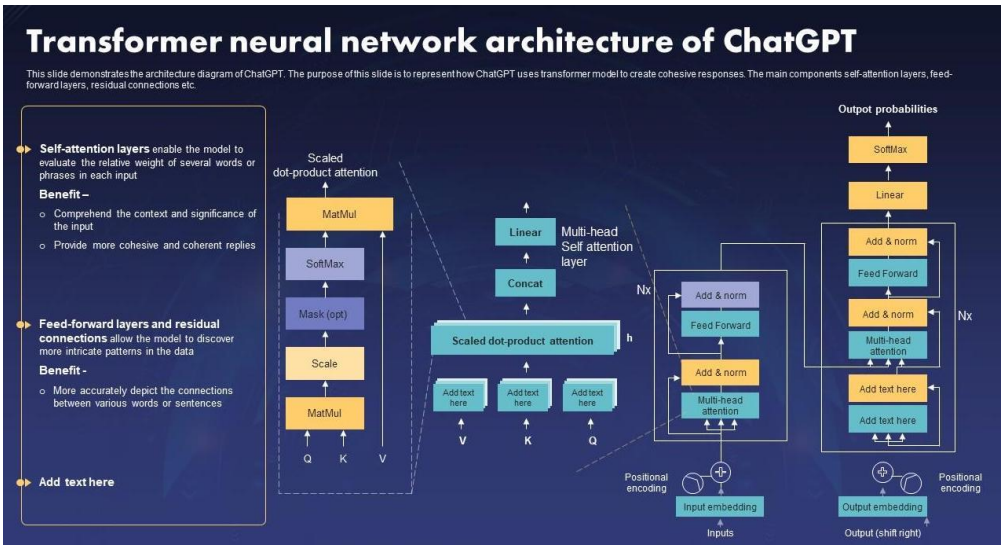

Fig. 1 Transformer neural network architecture of ChatGPT [5].

The GPT training process entails three systematic steps to optimize the model for specific tasks or domains (Fig. 2). First, demonstration data representing suitable task responses are collected, allowing the model to understand the target domain and train a supervised policy. Second, comparison data comprising ranked responses to various prompts are acquired and used to develop a reward model, quantifying response quality. Lastly, the Proximal Policy Optimization (PPO) reinforcement learning algorithm optimizes the model's policy against the reward model, refining its response generation capabilities based on reward signals. Consequently, the GPT training process combines collecting demonstration data, developing a reward model via comparison data, and optimizing the policy using PPO reinforcement learning to achieve enhanced task-specific performance.

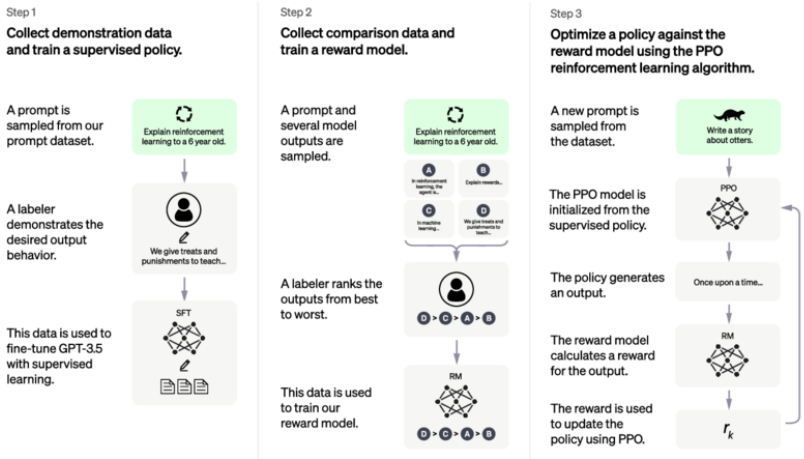

Fig. 2 ChatGPT training process [6].

2.2. Song Scoring

2.2.1. Score-Informed Networks

Huang et al. [7] employed three distinct deep neural network models to score musical performances. The first model, a Convolutional Neural Network (i.e. SIConvNet), directly predicts evaluation grades using aligned pitch contours and sheet music as simple time-sequential inputs. The second model, the Joint Embedding Model (i.e. JointEmbedNet), learns the joint latent space of pitch contours and sheet music, using cosine similarity to predict the evaluation grades. The third model, the Distance Matrix-based Convolutional Neural Network (DistMatNet), leverages patterns in the distance matrix between pitch contours and musical scores to predict the evaluation grade assigned by instructors.

For SIConvNet and JointEmbedNet, an N×2 matrix is created by stacking aligned pitch contours and sheet music, where N represents the sequence length of pitch contours. DistMatNet's input representation consists of a matrix of pairwise wrapped distances between pitch contours and MIDI pitch sequences. Concerning the model architecture, SIConvNet utilizes a simple 4-layer CNN to directly predict evaluation grades. JointEmbedNet employs two separate encoder networks to project sheet music and pitch contours into a joint latent space and uses the similarity between the two embeddings to predict evaluation scores. DistMatNet adopts a residual CNN architecture with the input being the distance matrix, enabling it to learn pitch differences to assess performances.

2.2.2. An Automated Measurement of Similarity

Tchernichovski et al. [8] proposed an automated program for measuring song similarity in order to assess the accuracy of vocal imitations. This program leverages contemporary spectral analysis techniques to characterize a song's acoustic structure, producing high-quality spectrograms and simplifying them into a set of basic acoustic features. Based on these attributes, the program can automatically detect similarities between songs.

Initially, the program calculates rhythm-synchronized spectrograms for reference and test songs, and computes three distinct characteristics at three separate tempos, including Wiener entropy, spectral continuity, and pitch. To address issues potentially arising from erroneous rhythm levels in beat tracking algorithms, the program calculates the aforementioned features across three different rhythm levels. In order to prevent certain test songs ("impostor" songs) from receiving high scores across all feature computations, the program employs random reference songs to calculate features and normalizes them by taking the mean and standard deviation.

The program evaluates song similarity by computing Euclidean distances between instructor and student songs at both small (7-millisecond time window) and large (50-millisecond time interval) scales, converting these distances into p-values. By setting similarity thresholds, the program can determine which song portions are similar. The program's performance is tested by comparing instructor and student songs, as well as randomly paired songs. The program's results exhibit a high correlation with the visual ratings of human observers, indicating its effectiveness in evaluating vocal similarity.

2.3. Automatically Generating Exercises

2.3.1. MEAWS

Percival introduced a Computer-Assisted Music Instruction and Teaching (CAMIT) system called Musician Evaluation and Audition for Strings (MEAWS) [9], aimed at helping music students enhance their sense of rhythm and violin intonation. Researchers utilized the GPT architecture and Constraint Satisfaction Programming (CSP) to generate exercises of varying difficulty levels, with constraints that include rhythmic complexity, note durations, and the allowance of rests.

Audio detection encompasses rhythm detection, segmenting audio into frames, computing the Root Mean Square (RMS) amplitude for each frame, and employing a threshold to detect rhythm onset points. Also, pitch detection uses a modified YIN algorithm in the spectral domain to detect pitch (fundamental frequency f0) and convert it into MIDI pitch values.

MEAWS offers visualization feedback for rhythm exercises by comparing expected beats to detected beats and using background colors to indicate potential errors. In intonation exercises, it displays pitch deviations with colored bars, signifying whether a note is sharp, flat, or fluctuating. This tool effectively helps music students enhance their skills and continuously evolves through user feedback and research.

2.3.2. PASSAROLA

The core methodology behind the PASSAROLA system for generating advanced music exercises is based on a technical framework that combines ABC musical notation and a Domain-Specific

Language (DSL) [10]. ABC notation, as a textual representation, allows users to describe melodies, chords, and other musical elements in a concise textual format. The PASSAROLA system can parse these descriptions and convert them into visualized sheet music and audible audio files.

The system's principled model includes several key technological components: First, it uses reusable templates and rich data types to define the structure and content of music exercises. These templates and type systems enable users without a computer science background to create complex music exercises easily. Second, the system employs a LaTeX-based notation and syntax to provide users with a powerful document writing environment for constructing exercise questions, solutions, and answer texts.

Another core technology of PASSAROLA is the dynamic programming approach, which supports the reuse of subproblems, making the exercise generation process more efficient. Through this method, the system can create exercises with step-by-step solutions that help students grasp music theory concepts and practical skills. Additionally, the system supports the use of external calculators and tools, such as Maxima and Perl, which can perform complex music theory and computational tasks, thus expanding the system's capabilities.

2.4. Music Generation

2.4.1. ChatMusician

ChatMusician, a Large Language Model (LLM), centers on the LLaMA2 model to understand and generate music through continuous pre-training and fine-tuning processes. Yuan et al. created the MusicPile corpus [1], a dataset specifically designed to enhance LLM's musical capabilities, encompassing musical knowledge, music theory Q&A, music summarization, mathematics, and code data. The MusicPile corpus includes not only general text data but also music-related instructions and chat data, as well as music metadata scraped from YouTube and music knowledge Q&A pairs generated by GPT-4.

In the music generation, ChatMusician can create conditional music compositions based on chords, melodies, motives, musical forms, and styles, while understanding and extracting motives and forms from musical pieces. Additionally, ChatMusician demonstrates its music understanding capability in the MusicTheoryBench benchmark test, a comprehensive test containing music knowledge and music reasoning problems to assess LLM's understanding of music.

To evaluate the quality of music generated by ChatMusician, researchers employ human assessments and specific quantitative metrics, such as phrase-level repetitiveness and parsing success rate. Furthermore, the average percentile score is used to measure the model's controllability. In terms of language capabilities, ChatMusician scores higher than baseline models in the MMLU test, indicating that its enhancements in music understanding and generation do not impair its general language processing abilities.

2.4.2. SongComposer

SongComposer is an innovative Large Language Model (LLM) designed specifically for songwriting [11]. It utilizes a symbolic song representation method that converts melodies and lyrics into a format comprehensible to LLMs. By incorporating musical notation as new vocabulary items in the LLM lexicon rather than using raw audio signals, this approach enhances encoding efficiency and flexibility.

During the pre-training phase, SongComposer is trained on extensive corpora of pure lyrics and pure melodies through a next-token prediction task. For melody data, pitch-transformation techniques are employed to augment the dataset. To ensure precise alignment between lyrics and melodies, researchers proposed a tuple data format, where each tuple represents a discrete music unit, which could be a lyric, melody, or a lyric-melody pair. This format separates different tuples with vertical bars and uses the sequences of tuples as LLM input.

To effectively process duration information in melodies, a logarithmic encoding scheme is adopted, converting continuous duration ranges into a set of discrete values. Simultaneously, for pitch processing, 120 unique music tokens representing the 12 pitch classes and 10 octave scales are introduced by expanding the vocabulary. Through these steps, SongComposer excels in multiple tasks, including lyrics to melody translation, melody to lyrics translation, song continuation, and song generation from textual descriptions, surpassing state-of-the-art LLMs in performance.

3. Discussion

3.1. Interpretability

The deployment of ChatGPT in music education, specifically in song scoring tasks, faces the challenge of limited interpretability. As an artificial intelligence model, ChatGPT frequently functions as a black-box system, meaning that its outputs may not provide a clear understanding of the final score assignment. This opacity affects both educators and learners in the music education domain. Discerning the reasons behind specific scores is essential for a meaningful learning experience and subsequent improvement for both teachers and students. This lack of transparent decision-making could lead to diminished trust in AI-generated scores. Furthermore, it hinders educators' ability to offer valuable feedback, as comprehending the basis of scores is crucial for addressing weaknesses and refining techniques.

Future research and development should investigate methods to enhance interpretability in ChatGPT's song scoring applications. One potential solution involves incorporating domain knowledge and expert systems during the model training process to increase system transparency. Another possibility is utilizing techniques such as Layer-wise Relevance Propagation (LRP) or Local Interpretable Model-agnostic Explanations (LIME) to shed light on the model's decision-making process. By improving interpretability, these challenges can be tackled, unlocking the full potential of ChatGPT within music education.

3.2. Algorithmic Improvement

The application of AI technology in music education, particularly for automatically generating exercises, is accompanied by several challenges. While most current technologies can handle simple answer types, they face limitations in generating more complex tasks.

Future developments could focus on extending system capabilities to accommodate a wider variety of exercises, integrating complex objects, and developing algorithms to improve the system's adaptability and flexibility. Moreover, emphasizing the effective use of computer technology to reinforce music learning is essential. Developing tools that provide accurate feedback and guidance is necessary. Some advanced AI methods can be further considered [12, 13]. Integrating audio analysis and visualization technologies, such as graphical interfaces that display rhythm and pitch variations, plays a crucial role in enhancing learning efficiency. Upcoming work should prioritize crafting more interactive and personalized learning tools that offer customized exercises and feedback based on students' progress and needs.

3.2.1. Applicability

In applying GPT models to music education for music generation, certain limitations are encountered, such as low adaptability and difficulties in generalizing across various styles and genres. These drawbacks impact the effectiveness of AI-generated compositions in meeting the diverse needs of music education.

Low adaptability constitutes a challenge, as GPT models are generally pretrained on large datasets and fine-tuned for specific tasks, while their inherent adaptability to various music styles remains limited. These models may struggle to accurately generate compositions in underrepresented genres or capture stylistic subtleties across different musical traditions. Utilizing domain adaptation

techniques can serve as a potential solution, fine-tuning the GPT models based on specific genres, styles, or user preferences. This approach leads to improved adaptability, resulting in more satisfactory music generation outcomes.

Another challenge lies in the difficulty of generalization. GPT models may struggle to generate high-quality compositions that can generalize across a range of genres and styles, potentially yielding outputs that fail to capture the essence of certain styles or adhere to the compositional rules associated with particular genres. Employing transfer learning enables GPT models to build on knowledge garnered from a broader musical corpus, attaining proficiency in generating compositions across diverse styles. Implementing multi-task learning approaches supports GPT models in learning shared features among related music generation tasks, potentially enhancing their generalization capabilities. Incorporating music theory, compositional rules, and expert knowledge into the model can help guide the AI system towards generating more precise and diverse music that is consistent with specific traditions and styles.

4. Conclusion

The study provides a comprehensive review of advancements in artificial intelligence algorithms, with an emphasis on the applications of GPT models in the music education domain. The research methodology employed in this study primarily includes the following:Utilizing contemporary spectral analysis techniques to deconstruct songs, and apply three deep neural network models for scoring music performances, Generating music exercises of varying difficulty levels using a technical framework that combines CSP, ABC notation, and DSL, Leveraging LLM with a focus on the LLaMA2 model, and continuously refining its understanding and generation of music through a process of pre-training and fine-tuning.

This paper also highlights the challenges in merging AI with music education, including issues like low adaptability and transparency. It explores solutions like domain adaptation and multi-task learning to overcome these hurdles. The discussion sheds light on the field's current limitations and improvement areas, aiming to enhance AI's role in music education. Despite offering a thorough review of AI's integration into music education, the fast-paced development of AI might limit this study's scope. Future research should focus on innovative approaches, keeping up with advancements to enrich music education with dynamic, personalized learning experiences.

References

[1] Yuan R, Lin H, Wang Y, et al. ChatMusician: Understanding and Generating Music Intrinsically with LLM. Hong Kong University of Science and Technology, 2024.

[2] Nogales Pérez D. A Deep Learning Based Tool for Ear Training, 2023.

[3] Guo Y, Liu Y, Zhou T, Xu L, Zhang Q. An automatic music generation and evaluation method based on transfer learning. Xihua University, 2023.

[4] Imasato N, Miyazawa K, Duncan C, Nagai T. Using a Language Model to Generate Music in Its Symbolic Domain While Controlling Its Perceived Emotion. Osaka University, 2023.

[5] Slideteam https://www.slideteam.net/media/catalog/product/cache/1280x720/t/r/transformer_neural_network_architecture_of_chatgpt_slide01.jpg, 2024.

[6] Megahed FM, et al. How generative AI models such as ChatGPT can be (mis) used in SPC practice, education, and research? An exploratory study. Quality Engineering, 2024: 287-315.

[7] Huang J, et al. Score-Informed Networks For Music Performance Assessment, 2020.

[8] Tchernichovski O, Pesaran B. A procedure for an automated measurement of similarity, 2000.

[9] Percival GK. Computer-Assisted Musical Instrument Tutoring with Targeted Exercises, 2006.

[10] Almeida JJ, Araújo I, Brito I, Carvalho N, Machado GJ, Pereira RMS, Smirnov G. PASSAROLA: High-Order Exercise Generation System, 2013.

[11] Ding S, Liu Z, Dong X, Zhang P, Qian R, He C, Lin D, Wang J. SongComposer: A Large Language Model for Lyric and Melody Composition in Song Generation, 2024.

[12] Qiu Y, Hui Y, Zhao P, et al. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy, 2024: 130866.

[13] Luo A, Zhong L, Wang J, et al. Short-term stock correlation forecasting based on CNN-BiLSTM enhanced by attention mechanism. IEEE Access, 2024.