**Chapter 2: Research Design**

Primary Data

**Sampling Strategy**

In the study conducted on the utilization of attention-based transformer models for polyphonic music generation, the primary research population consisted of five industry experts. An industry expert was defined as someone who was working in the music generation industry. Industry experts were sampled to obtain valuable insights from their in-depth knowledge of music creation, enriching the research with nuanced perspectives. A credible and reliable point of view was offered, and a unique perspective on the generated music was provided, which might not have been noticed otherwise.

The experts were gathered from LinkedIn and were not connected on LinkedIn prior to the research to avoid any response bias. LinkedIn was used as the sample frame of choice because it provided a large, rich pool of experts due to its global popularity and widespread adoption among professionals from various industries and geographical locations. It was easily accessible, and due to its user-friendly interface, it was virtually effortless to select a sample.

Due to time constraints, the research adopted a non-probability sampling technique, specifically expert sampling. While the probability sampling method could have provided a representative sample for statistical generalization, non-probability sampling was chosen for its efficiency in gathering feedback within a short timeframe. Judgement sampling was selected for exploratory research, allowing for the hand-picking of individuals with relevant experience and proficiency to provide valuable insight on the generating music. Experts were chosen to ensure diversity in expertise, perspectives, and backgrounds, encompassing various job titles across different companies. The research remained relevant to the overall research objectives, as the experts’ offered opinions on the coherence of the music, the impact of the dataset on its quality, and its potential for monotony. Their finely-tuned ears detected subtle nuances and intricacies in the music, providing insightful and knowledgeable input while optimizing time and resource utilization.

**Primary Research Methodology**

The method employed in the primary research involved conducting five expert interviews. Experts were contacted through LinkedIn and were invited to participate in the study. The topic of interest, research objectives, and the music generated from the analysis were shared with the interviewees prior to the interviews. Each interview lasted approximately forty-five minutes, with an additional fifteen minutes allocated for summarizing initial thoughts after the interview. Consent for recording the interviews was obtained through email or LinkedIn messages before the interviews occurred.

The expert interviews were conducted using a qualitative approach. Depending on logistical concerns, the interviews were organized for a mutually convenient time and were performed in person in a quiet public setting, such as a cafe, over the phone, or via Zoom. The experts had the opportunity to share their knowledge and opinions during the interviews on both the music generated and its alignment with the research objectives and music generation in general, providing a better understanding of the topic. To ensure the relevance of the information collected from the interviews, open-ended questions were designed to elicit detailed and insightful responses from the experts, facilitating exploration and clarification of their perspectives.

The audio recordings of the interviews were transcribed verbatim, and the transcripts served as the primary source of data for analysis. To identify patterns, themes, and significant findings in the interview data, thematic analysis, content analysis, or other qualitative analytic techniques were employed. The reliability and validity of the results were ensured through the application of rigorous analysis procedures. Once the analysis was finished, it was compared to the findings from secondary research to determine whether our results aligned with what the experts had found and, if not, to understand the reasons for any disparities.

The preference for the chosen method of primary research stemmed from the need for more than mere subjective judgments of the generated music's quality. In order to make informed decisions and derive meaningful conclusions from the research findings, a valid and relevant opinion from a credible evaluator was essential. This need was particularly pronounced in the pursuit of the first objective, aimed at assessing the self-attention transformer's ability to capture long-term dependencies in polyphonic music.

The expert evaluated whether the newly generated music constituted a coherent composition that seamlessly integrated with the trained music, distinguishing between a well-structured piece and a mere sequence of random notes. An expert's perspective was crucial in identifying the strengths and weaknesses of the music, especially regarding the second objective, which focused on potential quality changes resulting from the use of a new dataset with the model. The expert's keen sense for what constitutes high-quality music enabled them to discern any significant alterations, considering metrics such as pitch, rhythm, melody, harmony, originality, and emotional impact.

Furthermore, the use of expert interviews offered flexibility in shaping the questions posed. If the interviewee raised intriguing points not previously considered, further exploration was possible. Likewise, when encountering unclear aspects, the interviewee could provide immediate elaboration. This advantage of the research method proved invaluable when discussing the third objective, which delved into the fine line between music repetition for structural, emphatic, and rhythmic purposes without becoming monotonous and uninteresting to the listener. The expert's insights allowed for a deeper exploration and assessment of whether this balance had been achieved.

When employing this research method, it is essential to acknowledge certain limitations. Firstly, the subjective nature of expert opinions introduced the possibility of biases, as feedback was influenced by their individual experiences, knowledge, and personal inclinations. Additionally, the findings might have lacked generalizability to the broader population, as a limited sample of five experts was interviewed rather than a larger or more representative group. Nevertheless, despite these limitations, the benefits offered by this method outweighed the drawbacks, leading to its implementation in the research.

**Problem Identification and Clarification**

This thesis aims to explore the capacity of self-attention mechanisms within transformer models to effectively capture long-term dependencies when generating polyphonic music. These mechanisms enable the modeling of intricate relationships between various musical elements across extended time intervals, resulting in compositions that exhibit greater complexity and coherence. Additionally, this research will examine the interplay between the size and diversity of training data and the quality of the generated musical output. Throughout the project, various techniques will be employed to mitigate or eliminate the production of monotonous or mundane musical pieces.

The field of music generation has witnessed remarkable advancements in recent years, largely driven by the application of deep learning techniques. One such approach is the attention-based Transformer model, originally introduced in the "Attention Is All You Need" paper (Vaswani et al., 2017). This model has shown great promise in natural language processing tasks and has been successfully adapted for music generation.

The Transformer model, originally designed for natural language processing, has demonstrated its ability to capture long-range dependencies and understand the hierarchical structure of sequences. This makes it a promising candidate for modeling the complex relationships between multiple notes in polyphonic music. By leveraging the self-attention mechanism, the Transformer can attend to relevant parts of the input sequence and generate coherent and harmonically-rich music compositions.

The application of attention-based Transformer models to polyphonic music generation opens up exciting opportunities for creating original and expressive compositions. Researchers and practitioners in the field are actively exploring novel approaches and techniques to further enhance the capabilities of these models. Studying and advancing the generation of polyphonic music with attention-based Transformer models unlocks new avenues for creative musical expression and contributes to the development of intelligent systems capable of composing complex and engaging music.

**Research Objectives**

* Problem Identification: Transformer models have shown great promise in generating polyphonic music, but it is unclear how well they can capture long-term dependencies in the music.

Problem Clarification: The ability to capture long-term dependencies is important in generating music that has a coherent structure and is musically pleasing.

Problem Formulation: How effective are self-attention mechanisms in transformer models for capturing long-term dependencies in polyphonic music?

Objective: To evaluate the effectiveness of self-attention mechanisms in transformer models for capturing long-term dependencies in polyphonic music.

* This objective aims to investigate the suitability of transformer models for music generation by evaluating their ability to capture long-term dependencies in polyphonic music. The objective will involve examining the effectiveness of self-attention mechanisms in identifying and encoding relationships between musical elements over longer time periods.
* Problem Identification: The impact of training data on the performance of attention-based transformer models for polyphonic music generation is not well understood.

Problem Clarification: It is unclear how the size and diversity of training data affect the quality of generated music by attention-based transformer models.

Problem Formulation: The objective is to evaluate the impact of training data on the performance of attention-based transformer models for polyphonic music generation, and how the size and diversity of training data affects the quality of generated music.

Objective: To determine the relationship between training data size and diversity, and the quality of generated music by attention-based transformer models for polyphonic music generation.

* The objective of this study is to investigate how the quality of generated music by attention-based transformer models for polyphonic music generation is affected by the size and diversity of the training data used to train these models. In other words, the study aims to determine whether the quantity and variety of the training data have an impact on the quality of the generated music. By evaluating this relationship, the study can provide insights into how to optimize the training data selection process to improve the performance of attention-based transformer models for polyphonic music generation.
* Problem Identification: Computer-generated music often suffers from repetitive patterns, which can make the music uninteresting and predictable.

Problem Clarification: Generating diverse and original music is important in creating music that is musically pleasing and engaging.

Problem Formulation: How well can attention-based transformer models generate diverse and original polyphonic music, and how effective are they at avoiding repetitive patterns and generating novel musical ideas?

Objective: To assess the ability of attention-based transformer models to generate diverse and original polyphonic music by examining their ability to avoid repetitive patterns and generate novel musical ideas.

* The objective of this research is to evaluate the diversity and originality of polyphonic music generated by attention-based transformer models. Specifically, the research aims to examine the model's ability to avoid repetitive patterns and generate novel musical ideas. By assessing the model's ability to generate diverse and original music, this research can contribute to the development of more advanced and creative machine learning models for music generation.

**Validity Type**

The two components of validity management are relevant and reliable.

Relevant: The thesis aims to evaluate and justify the use of attention-based transformer models for polyphonic music generation. Previously, neural networks were used with limited success due to the exploding gradient problem. However, since the introduction of Large Language Models, it is deemed relevant to view music generation as a language modeling task and to apply attention-based transformers to this domain. The study will assess the relevance of utilizing these models in the context of music generation and evaluate their performance.

Reliable: The thesis will assess the effectiveness of attention-based transformer models for capturing long-term dependencies in music. The reliability of the findings will depend on the robustness of the evaluation methodology and the validity of the results.

**Ethical and Legal Considerations**

**Primary Research**

Ethical considerations were central to the Data Analysis project. Voluntary participation and informed consent from all five participating experts were required for the primary research. This was achieved by providing a clear explanation of the study's purpose, the nature of their involvement, and the utilization of their data via email/Linked In message. The option to seek clarification or pose questions beforehand through the same email chain was encouraged, with full respect for their decision to participate or withdraw from the study at any stage.

Trust and respect were of the utmost importance for all participants in the study. This meant showing up to the meeting early and being prepared to start with plenty of questions to fill the time slot agreed upon. Maintaining professionalism throughout the interview and ensuring their opinions and expertise were valued.

Ensuring the accuracy of result reporting is paramount. This process entails transcribing the recordings verbatim and extracting vital themes, quotes, and interesting insights while avoiding the inclusion of redundant or fabricated information.

When composing the thesis results, confidentiality and anonymity of the experts' responses will be maintained by assigning each of them an alias (e.g., interviewee A). All other personal information, including names, email addresses, Linked In profiles, phone numbers, etc., will be securely stored in private Linked In and/or email accounts with password protection and in a password-protected folder on a personal laptop. The findings will be presented in a manner that honours the experts' contributions and upholds the research's integrity. Themes and insights derived from all five interviews will be summarized, enhancing readability, and preserving the essence of each interviewee's input.

In conclusion, ethical principles take precedence in the thesis project. Voluntary participation and informed consent are essential, ensuring that all five experts involved in the interviews comprehend the study's purpose, their roles, and the utilization of their data. Trust and respect play a central role, emphasized through punctuality, preparedness, and genuine appreciation for the experts' opinions and expertise. Reporting and dissemination follow responsible practices, accurately conveying the results of data analysis through verbatim transcriptions and the extraction of key themes, quotes, and insights. To protect confidentiality and anonymity, experts receive aliases, and personal information is securely stored. Ultimately, the findings are presented in a format that honours the experts' contributions, delivering summarized themes and insights that preserve the research's integrity and engage readers effectively. These ethical considerations ensure a robust and respectful execution of the Primary Research in the thesis project.

**Secondary Research**

In the ever-evolving landscape of AI-generated music, a fundamental consideration lay in harmonious coexistence with established copyrights and the avoidance of infringement. Irish, UK, and USA copyright laws stated that 'copyright in a literary, dramatic, musical, or artistic work expired 70 years after the death of the author. After this period, the work was said to enter the public domain, allowing it to be used, modified, or republished by any person without fear of copyright infringement' (About Copyright, no date) (How copyright protects your work, no date) (Editor, 2022). This emancipated the creation, allowing it to be freely utilized, adapted, and republished without the looming specter of copyright infringement. As the exploration of AI-driven musical composition continued, it became pivotal to not only respect these legal frameworks but also to leverage open-source data discussed in the Data Collection section. This data aligned with copyright regulations and duly acknowledged the sources, thereby safeguarding intellectual property and artistic integrity.

When venturing into the realm of AI-generated music, it was crucial to remain vigilant about potential biases stemming from the training data and algorithms, which might result in the production of prejudiced or objectionable musical compositions. The landscape of music genres and their associated music theory was far from straightforward. Their development was influenced by a complex web of factors, often intersecting, and shaping the transformation of musical styles and practices. Music was profoundly shaped by elements such as geography, culture, religion, historical events, technological advancements, demographics, as well as the blending and fusion of various influences, to name just a few.

Critical aspects to be mindful of in the model's output were the potential presence of stereotypes and cultural biases in the music it generated. Stereotypes could emerge when the AI model, intentionally or unintentionally, replicated simplistic or biased notions about certain musical styles, genres, or cultures. For example, the model might falsely associate specific musical elements with cultural clichés, resulting in misrepresentations. These stereotypes could perpetuate cultural insensitivity and lead to feelings of hurt or disrespect among listeners.

Furthermore, cultural biases might also influence the AI-generated music, as they could manifest in the model's interpretation of various musical traditions and practices. Biased training data or human annotations might inadvertently introduce cultural biases into the system, leading to music that failed to authentically represent the rich diversity of musical heritage.

To mitigate these issues, it was imperative to choose the training dataset with meticulous care and a deep understanding of this nuanced reality. This involved not only diversifying the training data but also rigorously scrutinizing it for any pre-existing stereotypes and cultural biases. By doing so, it ensured that AI-generated music respected the rich tapestry of human culture and history while fostering creativity and innovation.

The most beautifully composed music possessed a unique quality that transcended mere ear-pleasing delight, having the power to stir deep emotions. These emotions could be both positive and negative. With this in mind, the responsibility that came with creating and sharing music was recognized. This responsibility extended to the well-being of the audience, and steps were taken to ensure that creations did not inadvertently cause discomfort or harm. By seeking feedback from a diverse group of listeners and making necessary adjustments, the endeavour was to craft music that touched hearts and minds without causing unintended distress. In this intricate dance between AI and art, music was created to resonate with people while respecting legal and cultural boundaries, ultimately offering a harmonious and inclusive musical journey.

Within the ever-shifting domain of AI-generated music, harmonious coexistence with copyright laws, vigilant avoidance of biases, and the creation of emotionally resonant compositions represented the pillars upon which the future of this art form stood. Navigating the intricate landscape of musical creation, the commitment to legal and ethical principles was steadfast, while also embracing the boundless possibilities that AI offered for creative expression. Upholding the values of respect, diversity, and emotional connection ensured that AI-generated music not only honoured the past but also paved the way for a harmonious and inclusive musical future. In this intricate fusion of technology and art, a symphony of potential was found, where AI and human creativity joined hands to compose a harmonious melody for generations to come.

**Chapter 3: Literature Review**

**RNN**

In the dynamic convergence of artificial intelligence and music, an extensive collection of research papers illuminates the multifaceted applications of neural networks, with a particular focus on Recurrent Neural Networks (RNNs), across diverse realms of music generation, transcription, and expressive performance.

The exploration begins with linguistic applications in (Lin et al., 2017), which meticulously evaluates a self-attention model across tasks such as author profiling, sentiment classification, and textual entailment. The identified significant performance gains underscore the prowess of self-attention mechanisms, setting the stage for a deeper dive into neural networks' applications.

In the realm of music generation and transcription, Recurrent Neural Networks (RNNs) stand as a cornerstone, as evidenced by several key research endeavours. (Sigtia, Benetos and Dixon, 2016) and (Goel, Vohra and Sahoo, 2014) delve into the application of RNNs, particularly Long Short-Term Memory (LSTM) networks, to capture the temporal intricacies of music. These papers showcase the pivotal role of RNNs in modeling sequences and temporal dependencies, a crucial aspect in both music transcription and polyphonic generation.

Additionally, the research paper (Roberts et al., 2019) introduces MusicVAE, a recurrent Variational Autoencoder (VAE) designed to enhance the modeling of sequences with long-term structure, particularly in musical notes. MusicVAE demonstrates superior performance in sampling, interpolation, and reconstruction compared to a "flat" baseline model, making it highly relevant and promising for advancing the field of music generation. The availability of the code and pre-trained models encourages further exploration and application of MusicVAE in both music generation and other types of sequential data.

Another significant contribution comes from the paper (Boulanger-Lewandowski, Bengio and Vincent, no date) In this paper, the authors address the modeling of symbolic sequences of polyphonic music using a general piano-roll representation. They introduce a probabilistic model that utilizes distribution estimators conditioned on a recurrent neural network to discover temporal dependencies in high-dimensional sequences. Their approach outperforms traditional models on various realistic datasets, particularly in the realm of polyphonic transcription, where it serves as a valuable symbolic prior, enhancing the accuracy of transcription.

Additionally, studies such as (Nikam, no date) and (Jaques et al., no date) extend the exploration of RNNs into music composition. The former introduces "Momentum LSTM" as an optimizer, underlining RNNs' ability to enhance model accuracy and broaden applications in music generation, while the latter combines supervised learning with reinforcement learning to refine LSTM-generated sequences, demonstrating the versatility of RNNs in generating pleasing and diverse musical compositions.

Moreover, the quest for expressive musical performance is addressed in (Oore et al., 2018), where an LSTM-based recurrent network is employed for direct performance generation. While highlighting the local strength of the system, this research acknowledges the ongoing challenge of imbuing RNN-generated music with a sophisticated long-term structure. Collectively, these studies underscore the fundamental role of RNNs in capturing temporal dependencies and shaping the landscape of AI-driven music applications, influencing transcription, composition, and expressive musical performance.

Venturing into genre-specific applications, (Hadjeres, Pachet and Nielsen, 2017) and (Sturm et al., 2016) delve into the realms of classical and traditional Celtic music, respectively. The former introduces DeepBach, a graphical model designed to generate polyphonic music in the style of Johann Sebastian Bach, demonstrating user control over constraints in music generation. The latter applies deep learning, specifically LSTM networks, to music transcription modeling with a focus on traditional Celtic music practice, highlighting the versatile application of RNNs across diverse musical genres.

Furthermore, the exploration of novel training techniques and fine-tuning methodologies is articulated in (Jaques et al., 2017). This paper not only enriches the understanding of RNNs but also showcases their adaptability in generating high-quality sequences while preserving original data knowledge and diversity across various domains, including musical melodies and computational molecular structures.

Collectively, these research endeavours form a rich tapestry that not only highlights the foundational significance of RNNs in capturing temporal dependencies within music but also explores their diverse applications across music transcription, composition, and genre-specific generation. In doing so, they collectively contribute to the evolving landscape of AI-driven music research.

**Self-Attention**

**Introduction to Attention-based Transformer Models**

Attention-based transformer models have been proposed as an alternative to RNNs for capturing long-term dependencies in music. Transformers are a type of deep learning model which has been a powerful tool for various natural language processing tasks, including language translation and text generation. (Vaswani et al., 2017) The Transformer model is a new kind of encoder-decoder model that uses self-attention to make sense of language sequences. This allows for parallel processing and thus makes it much faster than any other model with the same performance. They thus paved the way for modern language models (such as BERT (Devlin et al., 2019), GPT (Brown et al., 2020) and T5 (Raffel et al., 2020)) Attention-based transformer models are based on the self-attention mechanism, which allows the model to attend to different parts of the input sequence when making predictions. The self-attention mechanism in attention-based transformer models enables the model to weigh the importance of different positions in the input sequence, allowing it to focus on the most relevant information for a given task. This makes the model more flexible and adaptable to different input lengths and patterns compared to RNNs. This mechanism allows the model to capture long-term dependencies without the vanishing gradient problem seen in RNN models.

**Language Transformer models**

The landscape of language transformer models has undergone a profound evolution, as evidenced by an array of innovative research papers. Notably, (Parikh et al., 2016) and (Luong, Pham and Manning, 2015) delve into the realm of attention mechanisms, showcasing the versatility and efficacy of such mechanisms in various NLP tasks. While the former explores decomposable attention for natural language inference, the latter investigates global and local attention for machine translation. In a parallel exploration, (Peng et al., 2022) and (Stern et al., 2019) showcase the adaptability of transformer architectures. The former focuses on speaker verification with an emphasis on fine-tuning, while the latter introduces a novel insertion-based approach for sequence generation. Simultaneously, the paper (Shaw, Uszkoreit and Vaswani, 2018) contributes to this landscape by efficiently handling relative positions, enhancing performance in sequence tasks. The exploration of novel adaptations continues with (Wang, Lee and Chen, 2019) which introduces modifications to the bidirectional Transformer encoder, emphasizing the integration of tree structures into self-attention heads for improved interpretability and performance in NLP tasks. Concurrently, (Keskar et al., 2019) emerges as a standout contribution, pushing the boundaries of language generation with CTRL—a large-scale model designed for controllable text generation. This paper places explicit control over style, content, and task-specific behaviour at the forefront, contributing significantly to discussions on controllable language models. Expanding on the themes of versatility and adaptability, (Luong et al., 2016) explores the benefits of sharing encoders and decoders across different tasks, leveraging unsupervised data to enhance translation quality. Shifting focus to music recognition, (Reghunath and Rajan, 2022) demonstrates the effectiveness of combining transformer models in an ensemble voting scheme. Collectively, these research papers illustrate the multifaceted evolution of transformer models, emphasizing their adaptability, efficiency, controllability, and widespread applicability across the spectrum of challenges in natural language processing, inference, generation tasks, and even speaker verification.

**Challenges in Music Generation with Transformers**

However, Transformer models are not very practical to implement for music generation because of how computational expense it is to run. i.e. the square of the sequence length (Huang et al., 2018) proposed an algorithm that reduces their intermediate memory requirement to linear in the sequence length. This enabled them to demonstrate that a Transformer with modified relative attention mechanism can generate minute long compositions with compelling structure, generated continuations that coherently elaborate on a given motif, and in a seq2seq setup generate accompaniments conditioned on melodies.

However, even though transformers have a potential of learning longer-term dependency, they are also limited by a fixed-length context in the setting of language modelling. (Dai et al., 2019) proposed a novel neural architecture Transformer-XL that enables learning dependency beyond a fixed length without disrupting temporal coherence. It consists of a segment-level recurrence mechanism and a novel positional encoding scheme. Their method not only enables capturing longer-term dependency, but also resolves the context fragmentation problem. As a result, Transformer-XL learns dependency that is 80% longer than RNNs and 450% longer than vanilla Transformers A.K.A the original transformer by (Vaswani et al., 2017), achieves better performance on both short and long sequences, and is up to 1,800+ times faster than vanilla Transformers during evaluation. Leveraging off of this discovery (Donahue et al., 2019) used transfer learning procedure to generate video game sound synthesis chip music which is multi-instrumental music by pre-training on a widely used large-scale dataset called Lakh MIDI and using Transformer-XL. Showing that they improve results both quantitatively and qualitatively by pertaining on a cross-domain dataset. As well as generating both chiptunes from scratch and collaborating with human composers.

**Advancements in Multi-Sequence Music Generation**

(Wu, Wang and Lei, 2020) built on this idea of Transformer-XL and instead of a single sequence-based Transformer-XL they generated music with multiple sequence of time-valued notes. They aimed to address two challenges: computing notes with the same value but different tempos, and the model's limitation in separately learning music aspects like harmony and rhythm due to the use of a single sequence. However, by solving these issues it is adding more complexity to the transformer which can lead to longer training times and higher computational requirements. Similarly, (Ens and Pasquier, 2020) is a model that explores conditional multi-track music generation using the Transformer architecture. Although it is limited by its ability to only generate music for a fixed number of bars it has created a space to explore diverse musical styles and the creation of complex and engaging multi-track music. Building upon this research landscape, the paper by (Valenti, Berti and Bacciu, 2021) introduces Calliope, a Transformer-based autoencoder specifically designed for efficient modeling of multi-track polyphonic music. Calliope not only addresses the challenges of applying deep learning to music modeling but also showcases notable advancements in musical sequence reconstruction and generation, particularly for extended sequences. The promising results and outlined future directions, such as expanding to handle more tracks and incorporating controllable musical properties, position Calliope as a significant contribution to the evolving field.

This exploration resonates with the theme of multi-instrumental music generation tackled in (Li, 2021) In particular, this study focuses on the application of the attention-based Transformer neural network architecture in the domain, along with exploring different music representations, and the application of transfer learning in the context of music generation. The study highlights the Transformer model's ability to learn multi-instrumental chiptune music, indicating its potential extension to similar music domains. It introduces a novel cross-domain pre-training method, demonstrating task improvement with larger datasets. Additionally, the creation of an event-based representation tailored for multi-instrumental music enhances the overall research landscape.

Expanding upon the application of attention mechanisms in multi-instrumental music generation, (Hsiao et al., 2021) paper introduces a novel Transformer variant that processes multiple consecutive tokens simultaneously, utilizing token types for input and output customization. The model achieves efficient sequence compression through integrated token embeddings, displaying exceptional performance in music modeling. It generates full-song piano compositions of comparable quality to a Transformer-XL based model but with significantly reduced training and inference times.

**Prominent Models: MuseNet and Pop Music Transformer**

(MuseNet, 2019) developed by OpenAI, is a prominent AI model that employs a variant of the Transformer architecture to generate diverse and original musical compositions. Trained on a vast dataset of MIDI files, MuseNet demonstrates the ability to produce coherent and stylistically diverse compositions in various genres, offering potential applications in music production, creative inspiration, and entertainment. It only drawback is that it lacks a deep understanding of musical theory and context, sometimes resulting in compositions that may sound musically inconsistent or unfamiliar.

Contrastingly, (Huang and Yang, 2020) aim to focus on just Pop piano music and build a Pop Music Transformer. The paper highlights the effectiveness of the Transformer model for generating expressive classical piano performances, while proposing improvements in data representation to enhance music modeling. By incorporating a metrical structure in the input data, the Pop Music Transformer is developed to generate Pop piano music with improved rhythmic and harmonic structures. These advancements showcase the Transformer's ability to learn abstractions and generate coherent compositions without relying heavily on human-imposed constraints or domain knowledge. The only issue with this approach is that it makes it less versatile for generating music outside the pop genre.

Likewise, the work done on music generation becomes null and void if it is not interacted with by musicians. Generative algorithms are still not widely used by artists due to the limited control they offer, prohibitive inference times or the lack of integration within musicians’ workflows.

The exploration of Sparse Transformers (Child et al., 2019) sets the stage for handling extreme-length sequences by modifying the Transformer model to reduce time and memory complexity. This foundational paper introduces techniques that impact sequence modeling, creating a bridge to subsequent works. Building upon the theme of modeling long-range sequences, the Compressive Transformer (Rae et al., 2019) is introduced, emphasizing its effectiveness in language modeling benchmarks and memory tasks. The insights gained from this paper contribute to the broader understanding of sequence modeling, laying the groundwork for subsequent research. In a related vein, (Cope, 1989) introduces a Transformer autoencoder for conditional music generation, leveraging autoregressive Transformer encoders and decoders. Through the combination of global style representations and temporally distributed embeddings, the model enhances control over various music generation aspects. Demonstrated effectiveness on diverse tasks using datasets like MAESTRO and YouTube, it achieves improved log-likelihood and mean listening scores compared to baseline models. Another noteworthy contribution emerges from (Guo et al., 2023). This study introduces the MT-GPT-2 model, a music generation model based on a text-like representation of music that includes pitch, rhythm, and pauses. The model is evaluated using a novel method called MEM, combining mathematical statistics and music theory, and it outperforms other existing models in generating music closer to real compositions. Together, these research endeavours represent significant strides in sequence modeling, particularly in the domains of handling extreme-length sequences and advancing the state of the art in music generation and evaluation.

**Human-Machine Interaction in Music Generation**

(Hadjeres and Crestel, 2021) tackles this by presenting the Piano Inpainting Application (PIA), a generative model focused on “inpainting” piano performances, as they believe that elementary operation (restoring missing parts of a piano performance) encourages human-machine interaction and opens up new ways to approach music composition. It allows musicians to smoothly generate or modify any MIDI clip using PIA within a widely used professional Digital Audio Workstation. The transformer used in this approach is the Linear Transformer by (Katharopoulos et al., 2020) which is performs similar to the original vanilla Transformer but can be up to 4000x faster on autoregressive prediction of very long sequences.

**Style Representation and Emotion in Music Generation**

(Choi et al., 2020) explores the use of Transformer autoencoders to encode and decode musical style representations. By leveraging the Transformer architecture, the model can learn to capture and reconstruct the underlying stylistic elements of music, allowing for style manipulation and generation. The approach shows promise in enabling fine-grained control over musical style and opens up possibilities for style transfer and composition. As well as music being generated in a specific style a crucial element of make great music is when it evokes an emotion from the listener. (Makris, Agres and Herremans, 2021) introduces a novel approach to generating lead sheets that incorporates high-level musical characteristics, specifically valence (the positivity or negativity of the perceived emotion), for control over the generated output. By using pre-defined mood tags and a conditional sequence-to-sequence framework, the authors demonstrate the ability to generate lead sheets in a controllable manner, achieving distributions of musical attributes similar to the training data and effective control over the valence of the generated chord progressions. This human-like element to the music being generated brings us closer to creating emotionally expressive and authentic compositions that resonate with listeners on a deeper level, bridging the gap between machine-generated music and the artistry of human musicians.

**Transformer-GANs: Combining Strengths**

Finally, the model that appears to yield the most favourable results in comparison to the Vanilla Transformer and Music Transformer is the Transformer-GANs. Drawing on a diverse range of methodologies explored above, it emerges as a compelling solution, offering superior outcomes in terms of music generation by leveraging the strengths of both the Transformer architecture and Generative Adversarial Networks. (Muhamed et al., 2021), (Muhamed et al., 2021), (Neves, Fornari and Florindo, 2022) and (Symbolic Music Generation with Transformer-GANs) use this approach with great success. They propose a framework that combines the Transformer architecture and Generative Adversarial Networks (GANs) for generating music with specific sentiments. The authors introduce a sentiment encoder to condition the generation process, allowing control over the emotional content of the generated music. Experimental results demonstrate the effectiveness of their approach in producing music that conveys desired sentiments while maintaining musical coherence and quality.

**Drawbacks of Transformer-GANs**

However, a huge drawback and reason why we would not use it in our strategy is Transformer-GANs is very complex and computational costly when training the model. Transformer-GANs combine the Transformer architecture with Generative Adversarial Networks, which both require significant computational resources and training time. This can pose challenges in terms of scalability and practicality, especially when dealing with large datasets or real-time music generation scenarios. Additionally, ensuring the stability and convergence of the GAN training process can be challenging, requiring careful tuning and experimentation.

**Dataset**

In the realm of music generation, the availability and quality of training datasets play a pivotal role in shaping the creative potential and output of generative models. This literature review explores the landscape of music datasets, for example symbolic datasets, audio datasets, lyrics and text datasets, music notation datasets, genre-specific datasets, emotional or mood-based datasets to name a few. As researchers and artists alike seek to push the boundaries of what is possible in automated music creation, the choice and characteristics of the underlying data become of paramount importance.

**Music**

**Repetitive Pattern**

**Chords**

**Sampling**

**References**

Boulanger-Lewandowski, N., Bengio, Y. and Vincent, P. (no date) ‘Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription’.

Brown, T.B. et al. (2020) ‘Language Models are Few-Shot Learners’. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.14165.

Child, R. et al. (2019) ‘Generating Long Sequences with Sparse Transformers’. arXiv. Available at: https://doi.org/10.48550/arXiv.1904.10509.

Choi, K. et al. (2020) ‘Encoding Musical Style with Transformer Autoencoders’. arXiv. Available at: https://doi.org/10.48550/arXiv.1912.05537.

Cope, D. (1989) ‘Experiments in musical intelligence (EMI): Non‐linear linguistic‐based composition’, Interface, 18(1–2), pp. 117–139. Available at: https://doi.org/10.1080/09298218908570541.

Dai, Z. et al. (2019) ‘Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context’. arXiv. Available at: https://doi.org/10.48550/arXiv.1901.02860.

Devlin, J. et al. (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. arXiv. Available at: https://doi.org/10.48550/arXiv.1810.04805.

Donahue, C. et al. (2019) ‘LakhNES: Improving multi-instrumental music generation with cross-domain pre-training’. arXiv. Available at: https://doi.org/10.48550/arXiv.1907.04868.

Ens, J. and Pasquier, P. (2020) ‘MMM : Exploring Conditional Multi-Track Music Generation with the Transformer’. arXiv. Available at: https://doi.org/10.48550/arXiv.2008.06048.

Goel, K., Vohra, R. and Sahoo, J.K. (2014) ‘Polyphonic Music Generation by Modeling Temporal Dependencies Using a RNN-DBN’, in, pp. 217–224. Available at: https://doi.org/10.1007/978-3-319-11179-7\_28.

Guo, Y. et al. (2023) ‘An automatic music generation and evaluation method based on transfer learning’, PLOS ONE, 18(5), p. e0283103. Available at: https://doi.org/10.1371/journal.pone.0283103.

Hadjeres, G. and Crestel, L. (2021) ‘The Piano Inpainting Application’. arXiv. Available at: https://doi.org/10.48550/arXiv.2107.05944.

Hadjeres, G., Pachet, F. and Nielsen, F. (2017) ‘DeepBach: a Steerable Model for Bach Chorales Generation’. arXiv. Available at: https://doi.org/10.48550/arXiv.1612.01010.

Hsiao, W.-Y. et al. (2021) ‘Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs’. arXiv. Available at: http://arxiv.org/abs/2101.02402 (Accessed: 18 August 2023).

Huang, C.-Z.A. et al. (2018) ‘Music Transformer’. arXiv. Available at: http://arxiv.org/abs/1809.04281 (Accessed: 31 March 2023).

Huang, Y.-S. and Yang, Y.-H. (2020) ‘Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions’. arXiv. Available at: http://arxiv.org/abs/2002.00212 (Accessed: 15 May 2023).

Jaques, N. et al. (2017) ‘Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control’. arXiv. Available at: http://arxiv.org/abs/1611.02796 (Accessed: 31 March 2023).

Jaques, N. et al. (no date) ‘Generating Music by Fine-Tuning Recurrent Neural Networks with Reinforcement Learning’.

Katharopoulos, A. et al. (2020) ‘Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention’. arXiv. Available at: https://doi.org/10.48550/arXiv.2006.16236.

Keskar, N.S. et al. (2019) ‘CTRL: A Conditional Transformer Language Model for Controllable Generation’. arXiv. Available at: https://doi.org/10.48550/arXiv.1909.05858.

Li, Y. (2021) Application of Attention Mechanism on Multi-Instrumental Music Generation. UC San Diego. Available at: https://escholarship.org/uc/item/0qw6q9sn (Accessed: 15 May 2023).

Lin, Z. et al. (2017) ‘A Structured Self-attentive Sentence Embedding’. arXiv. Available at: https://doi.org/10.48550/arXiv.1703.03130.

Luong, M.-T. et al. (2016) ‘Multi-task Sequence to Sequence Learning’. arXiv. Available at: https://doi.org/10.48550/arXiv.1511.06114.

Luong, M.-T., Pham, H. and Manning, C.D. (2015) ‘Effective Approaches to Attention-based Neural Machine Translation’. arXiv. Available at: https://doi.org/10.48550/arXiv.1508.04025.

Makris, D., Agres, K.R. and Herremans, D. (2021) ‘Generating Lead Sheets with Affect: A Novel Conditional seq2seq Framework’. arXiv. Available at: https://doi.org/10.48550/arXiv.2104.13056.

Muhamed, A. et al. (2021) ‘Symbolic Music Generation with Transformer-GANs’, Proceedings of the AAAI Conference on Artificial Intelligence, 35(1), pp. 408–417. Available at: https://doi.org/10.1609/aaai.v35i1.16117.

MuseNet (2019). Available at: https://openai.com/research/musenet (Accessed: 15 May 2023).

Neves, P., Fornari, J. and Florindo, J. (2022) ‘Generating music with sentiment using Transformer-GANs’. arXiv. Available at: http://arxiv.org/abs/2212.11134 (Accessed: 14 May 2023).

Nikam, S. (no date) ‘Composition and production of music using Momentum LSTM’.

Oore, S. et al. (2018) ‘This Time with Feeling: Learning Expressive Musical Performance’. arXiv. Available at: http://arxiv.org/abs/1808.03715 (Accessed: 15 May 2023).

Parikh, A.P. et al. (2016) ‘A Decomposable Attention Model for Natural Language Inference’. arXiv. Available at: http://arxiv.org/abs/1606.01933 (Accessed: 15 May 2023).

Peng, J. et al. (2022) ‘An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification’. arXiv. Available at: http://arxiv.org/abs/2210.01273 (Accessed: 15 May 2023).

Rae, J.W. et al. (2019) ‘Compressive Transformers for Long-Range Sequence Modelling’. arXiv. Available at: https://doi.org/10.48550/arXiv.1911.05507.

Raffel, C. et al. (2020) ‘Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer’. arXiv. Available at: https://doi.org/10.48550/arXiv.1910.10683.

Reghunath, L.C. and Rajan, R. (2022) ‘Transformer-based ensemble method for multiple predominant instruments recognition in polyphonic music’, EURASIP Journal on Audio, Speech, and Music Processing, 2022(1), p. 11. Available at: https://doi.org/10.1186/s13636-022-00245-8.

Roberts, A. et al. (2019) ‘A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music’. arXiv. Available at: http://arxiv.org/abs/1803.05428 (Accessed: 15 May 2023).

Shaw, P., Uszkoreit, J. and Vaswani, A. (2018) ‘Self-Attention with Relative Position Representations’. arXiv. Available at: http://arxiv.org/abs/1803.02155 (Accessed: 15 May 2023).

Sigtia, S., Benetos, E. and Dixon, S. (2016) ‘An End-to-End Neural Network for Polyphonic Piano Music Transcription’. arXiv. Available at: http://arxiv.org/abs/1508.01774 (Accessed: 4 August 2023).

Stern, M. et al. (2019) ‘Insertion Transformer: Flexible Sequence Generation via Insertion Operations’, in Proceedings of the 36th International Conference on Machine Learning. International Conference on Machine Learning, PMLR, pp. 5976–5985. Available at: https://proceedings.mlr.press/v97/stern19a.html (Accessed: 15 May 2023).

Sturm, B.L. et al. (2016) ‘Music transcription modelling and composition using deep learning’. arXiv. Available at: https://doi.org/10.48550/arXiv.1604.08723.

Valenti, A., Berti, S. and Bacciu, D. (2021) ‘Calliope -- A Polyphonic Music Transformer’. arXiv. Available at: http://arxiv.org/abs/2107.05546 (Accessed: 4 August 2023).

Vaswani, A. et al. (2017) ‘Attention Is All You Need’. arXiv. Available at: http://arxiv.org/abs/1706.03762 (Accessed: 1 April 2023).

Wang, Y.-S., Lee, H.-Y. and Chen, Y.-N. (2019) ‘Tree Transformer: Integrating Tree Structures into Self-Attention’. arXiv. Available at: https://doi.org/10.48550/arXiv.1909.06639.

Wu, X., Wang, C. and Lei, Q. (2020) ‘Transformer-XL Based Music Generation with Multiple Sequences of Time-valued Notes’. arXiv. Available at: https://doi.org/10.48550/arXiv.2007.07244.