Generating Polyphonic Music with Attention-Based Transformer Model

Naomi Tunstead

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

February 2024

Supervisor: Taufique Ahmed

**Acknowledgement**

I would like to express my deepest gratitude to my supervisor, Taufique Ahmed, whose guidance, support, and expertise have been invaluable throughout the entirety of my master's journey. Your dedication to my academic and personal growth has played a pivotal role in shaping this thesis and my overall development as a researcher.

I am also profoundly thankful to my family and friends for their unwavering encouragement, understanding, and patience during this challenging yet rewarding endeavour. Your belief in me and constant support have been a constant source of strength and motivation.

Last but certainly not least, I extend my heartfelt thanks to God, whose grace, guidance, and blessings have sustained me through every trial and triumph. Without His divine intervention and unwavering presence, this accomplishment would not have been possible.

To all those who have contributed to this journey in ways both seen and unseen, I am deeply indebted and forever grateful.

Thank you.

**Abstract**

This thesis investigates attention-based transformer models for AI-generated polyphonic music. Three primary objectives are addressed: the impact of training data on model performance, the effectiveness of self-attention mechanisms, the mitigation of repetitive patterns, and the promotion of music diversity. Transformer models, surpassing traditional recurrent neural networks, exhibit potential in capturing long-term dependencies. The study emphasizes the significance of training data characteristics, self-attention mechanisms, and pattern diversity in enhancing music quality. Through human-centred methodologies, the research offers insights into music perception. Future research avenues include broadening participant demographics, exploring alternative model architectures, and further mitigating repetitive patterns to advance AI-generated music.

**Keywords:** Attention-based transformer models, Training data, Self-attention mechanisms, Repetitive patterns, Music diversity, Human-centred methodologies

**Table of Contents**

[Chapter 1: Introduction 1](#_Toc159600507)

[1.1 Motivation 1](#_Toc159600508)

[1.2 Background 1](#_Toc159600509)

[1.3 Research Problems and Research Questions 3](#_Toc159600510)

[Chapter 2: Literature Review 4](#_Toc159600511)

[2.1 RNN 4](#_Toc159600512)

[2.2 Self-Attention 6](#_Toc159600513)

[2.2.1 Introduction to Attention-based Transformer Models 6](#_Toc159600514)

[2.2.2 Language Transformer models 7](#_Toc159600515)

[2.2.3 Challenges in Music Generation with transformers 7](#_Toc159600516)

[2.2.4 Advancements in Multi-Sequence Music Generation 8](#_Toc159600517)

[2.2.5 Prominent Models: MuseNet and Pop Music Transformer 9](#_Toc159600518)

[2.2.6 Human-Machine Interaction in Music Generation 10](#_Toc159600519)

[2.2.7 Style Representation and Emotion in Music Generation 10](#_Toc159600520)

[2.2.8 Transformer-GANs: Combining Strengths 11](#_Toc159600521)

[2.2.9 Drawbacks of Transformer-GANs 11](#_Toc159600522)

[2.3 Repetitive Pattern 12](#_Toc159600523)

[2.4 Conclusion 13](#_Toc159600524)

[Chapter 3: Research Design and Methodology 15](#_Toc159600525)

[3.1 Dataset 15](#_Toc159600526)

[3.1.1 Primary Dataset 15](#_Toc159600527)

[3.1.2 Secondary Dataset 17](#_Toc159600528)

[3.2 Data Preprocessing 18](#_Toc159600529)

[3.2.1 MMM Tokenizer 18](#_Toc159600530)

[3.2.2 Training the Tokenizer 21](#_Toc159600531)

[3.3 Designing the Transformer Model 21](#_Toc159600532)

[3.3.1 Decoder Transformer Architecture 22](#_Toc159600533)

[3.4 Self-attention 23](#_Toc159600534)

[3.4.1 Multi-Head Attention 24](#_Toc159600535)

[3.5 Repetitive patterns 25](#_Toc159600536)

[3.6 Hyperparameters Optimization 26](#_Toc159600537)

[3.7 Model Evaluation Metrics 27](#_Toc159600538)

[Chapter 4: Results and Discussion 27](#_Toc159600539)

[Chapter 5: Ethical and Legal Considerations 39](#_Toc159600540)

[5.1 Primary Research 39](#_Toc159600541)

[5.2 Secondary Research 40](#_Toc159600542)

[Chapter 6: Conclusion 42](#_Toc159600543)

[6.1 Future Research 43](#_Toc159600544)

[References 45](#_Toc159600545)

[Appendix 50](#_Toc159600546)

[Appendix A: Github Link 50](#_Toc159600547)

[Appendix B: Open-Source Data Permissions 50](#_Toc159600548)

[Appendix C: Survey Link 50](#_Toc159600549)

[Appendix D:Interview Permissions 50](#_Toc159600550)

[Appendix E: Interview Transcripts 53](#_Toc159600551)

[Participant 1 53](#_Toc159600552)

[Participant 2 66](#_Toc159600553)

[Participant 3 89](#_Toc159600554)

[Participant 4 106](#_Toc159600555)

[Glossary 131](#_Toc159600556)

**List of Figures**

[Figure 1: The MultiTrack and BarFill representations are shown. The <bar> tokens correspond to complete bars, and the <track> tokens correspond to complete tracks. 20](#_Toc159536191)

[Figure 2: Original Transformer model architecture from Attention Is All You Need. The encoder is depicted on the right-hand side and the decoder on the left hand side (Vaswani et al., 2017) 23](#_Toc159536192)

[Figure 3: Musical Participants Evaluating Originality of Music Sample 28](#_Toc159536193)

[Figure 4: Non-Musical Participants Evaluating Originality of Music Sample 29](#_Toc159536194)

[Figure 5: Musical Participants Detecting Repetitive Patterns 29](#_Toc159536195)

[Figure 6: Non-Musical Participants Detecting Repetitive Patterns 30](#_Toc159536196)

[Figure 7: Musical Participants Evaluating Uniqueness of Music Ideas 30](#_Toc159536197)

[Figure 8: Non-Musical Participants Evaluating Uniqueness of Music Ideas 31](#_Toc159536198)

[Figure 9: Musical Participants Evaluating Perceived Musical Engagement Levels 32](#_Toc159536199)

[Figure 10: Non-Musical Participants Evaluating Perceived Musical Engagement Levels 32](#_Toc159536200)

[Figure 11: Music Participants Evaluating the Coherence of Music Sample 33](#_Toc159536201)

[Figure 12: Non-Music Participants Evaluating the Coherence of Music Sample 33](#_Toc159536202)

[Figure 13: Musical Participants Rating of Pleasure in Music Samples 34](#_Toc159536203)

[Figure 14: Non-Musical Participants Rating of Pleasure in Music Samples 34](#_Toc159536204)

[Figure 15: Musical Participants Evaluating the Quality of the Music Sample 35](#_Toc159536205)

[Figure 16: Non-Musical Participants Evaluating the Quality of the Music Sample 36](#_Toc159536206)

[Figure 17: Musical Participants Rating of Overall Music Sample Quality 36](#_Toc159536207)

[Figure 18: Non-Musical Participants Rating of Overall Music Sample Quality 37](#_Toc159536208)

[Figure 19: Musical Participants Perception of Music Flow and Coherence 37](#_Toc159536209)

[Figure 20: Non-Musical Participants Perception of Music Flow and Coherence 38](#_Toc159536210)

**List of Tables**

[Table 1: The hyperparameter values used in the GPT-2 model training 26](#_Toc159571403)

[Table 2: Glossary of Terms 128](#_Toc159571404)

# Chapter 1: Introduction

## 1.1 Motivation

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.

## 1.2 Background

The field of artificial intelligence (AI) and music creation has been extensively studied, offering deep insights into the complex relationship that exists between artistic expression and computational algorithms. Particularly noteworthy is the fundamental role played by recurrent neural networks (RNNs), notably Long Short-Term Memory (LSTM) networks. These networks excel in capturing temporal dependencies and modeling sequences in music generation, as demonstrated by various studies (Sigtia, Benetos and Dixon, 2016), (Goel, Vohra and Sahoo, 2014). Their contributions have significantly shaped the landscape of AI-driven music applications, influencing transcription, composition, and expressive musical performance (Hadjeres, Pachet and Nielsen, 2017), (Sturm *et al.*, 2016), (Oore *et al.*, 2018).

The emergence of attention-based transformer models presents a promising alternative to RNNs in capturing long-term dependencies in music (Parikh et al., 2016; Luong, Pham and Manning, 2015). These models leverage self-attention mechanisms to process input sequences in parallel, offering flexibility to different input lengths and patterns (Vaswani *et al.*, 2017). Despite challenges such as computational complexity and fixed-length context constraints, innovative solutions like modified attention mechanisms and novel neural architectures like Transformer-XL have been proposed to enhance transformer-based music generation (Shaw, Uszkoreit and Vaswani, 2018), (Dai *et al.*, 2019).

Furthermore, researchers are making strides in addressing challenges associated with tempo variations and the separate learning of musical elements such as harmony and rhythm (Wang *et al.*, 2020), (Ens and Pasquier, 2020a). Moreover, AI models like MuseNet and Pop Music Transformer show how effective transformer architectures are in creating diverse and authentic music across different genres (*MuseNet*, 2019), (Huang and Yang, 2020). Additionally, researchers are actively implementing techniques such as motif identification, theme-based conditioning, and sampling strategies like nucleus sampling and temperature control to alleviate issues of repetitive patterns and enhance the richness and diversity of generated music (Wang *et al.*, 2023), (Shih *et al.*, 2022), (Bjare, Lattner and Widmer, 2023).

However, research gaps in the field of artificial intelligence and music creation have included understanding the influence of training data size and diversity on the performance of attention-based transformer models for polyphonic music generation, assessing the effectiveness of self-attention mechanisms in capturing long-term dependencies in polyphonic music, and tackling the challenge of generating diverse and original polyphonic music while avoiding repetitive patterns. These gaps have indicated the need for further exploration and development to advance the capabilities of AI-driven music composition.

## 1.3 Research Problems and Research Questions

1. Research Problem Identification: the impact of training data on the performance of attention-based transformer models for polyphonic music generation is not well understood.
   1. Problem Clarification: it is unclear how the size and diversity of training data affect the quality of music generated by attention-based transformer models.
   2. 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.
   3. Objective: to determine the relationship between training data size and diversity and the quality of music generated by attention-based transformer models for polyphonic music generation.
2. Research 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.
   1. Problem Clarification: the ability to capture long-term dependencies is important in generating music that has a coherent structure and is musically pleasing.
   2. Problem Formulation: how effective are self-attention mechanisms in transformer models for capturing long-term dependencies in polyphonic music?
   3. Objective: to evaluate the effectiveness of self-attention mechanisms in transformer models for capturing long-term dependencies in polyphonic music.
3. Research Problem Identification: computer-generated music often suffers from repetitive patterns, which can make the music uninteresting and predictable.
   1. Problem Clarification: generating diverse and original music is important in creating music that is musically pleasing and engaging.
   2. 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?
   3. 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.

# Chapter 2: Literature Review

Research on artificial intelligence in music composition has produced a wealth of knowledge about attention-based transformer models as well as recurrent neural networks (RNNs). RNNs, especially LSTM networks, have been crucial in the modeling of temporal dependencies and sequences, as demonstrated by their many uses in music production. Advances in the field, such as MusicVAE and new approaches to polyphonic music, demonstrate how flexible RNNs are in producing high-quality sequences in a variety of contexts.

However, with recent developments, attention-based transformer models have become powerful rivals to replace RNNs in the long-term dependency capture task of music production. Transformer models process input sequences in parallel by utilizing self-attention mechanisms, which improves their flexibility to different input lengths and speeds up computation. Due to issues with computational complexity and fixed-length context constraints in transformer music creation, innovations such as GPT-2 have been developed to solve longer-term dependencies without compromising temporal coherence (Radford *et al.*, no date). Prominent models such as Pop Music Transformer and MuseNet serve as examples of how versatile and emotive musical compositions may be produced with the transformer.

Last but not least, recent developments in deep learning models-driven music generation, especially with regard to Transformer architectures, have demonstrated encouraging outcomes in resolving concerns like monotonous patterns and enhancing diversity in generated music. Research such as the "Theme Transformer" and the "Motif Transformer" provide creative methods for identifying unique themes and thematic development, respectively. All things considered, these group efforts provide a noteworthy advancement in the comprehension and improvement of transformer-based methods for producing music, providing insightful information for further advancements in the area.

## 2.1 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 domains of music generation, transcription, and expressive performance.

The exploration begins with linguistic applications (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 capability of self-attention mechanisms, setting the stage for a deeper dive into neural networks' applications.

In the domain 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 domain of polyphonic transcription, where it serves as a valuable symbolic prior, enhancing the accuracy of transcription.

Additionally, studies such as (Nikam, no date), (*Performance RNN: Generating Music with Expressive Timing and Dynamics*, 2017), 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 domains 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.

## 2.2 Self-Attention

### 2.2.1 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 that 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.

### 2.2.2 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 domain 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 and 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) emerged 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) explore the benefits of sharing encoders and decoders across different tasks, leveraging unsupervised data to enhance translation quality. Shifting the focus to music recognition (Reghunath and Rajan, 2022) demonstrate 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.

### 2.2.3 Challenges in Music Generation with Transformers

However, transformer models are not very practical to implement for music generation because of how much 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 a modified relative attention mechanism can generate minute-long compositions with compelling structure, generate continuations that coherently elaborate on a given motif, and, in a seq2seq setup, generate accompaniments conditioned on melodies.

However, even though transformers have the potential to learn longer-term dependency, they are also limited by a fixed-length context in the setting of language modeling. (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 this discovery (Donahue et al., 2019) a 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 to a cross-domain dataset. As well as generating both chiptunes from scratch and collaborating with human composers.

### 2.2.4 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 sequences 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, solving these issues adds more complexity to the transformer, which can lead to longer training times and higher computational requirements. Similarly, (Ens and Pasquier, 2020a) is a model that explores conditional multi-track music generation using the Transformer architecture. Although it is limited by its ability to generate music for a fixed number of bars only, 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.

### 2.2.5 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. Its 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.

In contrast, (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, 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.*, 2023a). 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 state of the art in music generation and evaluation.

### 2.2.6 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 performs similarly to the original vanilla Transformer but can be up to 4000x faster on autoregressive predictions of very long sequences.

### 2.2.7 Style Representation and Emotion in Music Generation

(Imasato *et al.*, 2023) presents a GPT-2-based transformer model for symbolic music generation, facilitating collaborative composition between humans and computers. With 939 participants, the study demonstrates promising results in controlling music generation based on desired emotions. This underscores the potential of AI in enabling collaborative music composition between humans and computers, highlighting the importance of user agency in creative tasks. Additionally, (Choi et al., 2020) explore 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 making 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.

### 2.2.8 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- Generative Adversarial Networks (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.*, 2021a), (Muhamed *et al.*, 2021a), (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.

### 2.2.9 Drawbacks of Transformer-GANs

However, a huge drawback and reason why the use of Transformer-GANs is not included in the strategy is that Transformer-GANs are very complex and computationally costly when training the model. Transformer-GANs combine the Transformer architecture with Generative Adversarial Networks, both of which 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.

## 2.3 Repetitive Pattern

As discussed previously, music generation has been a significant area of research in recent years, with advancements driven by deep learning models, particularly Transformer architectures. Several studies have focused on addressing issues such as repetitive patterns and enhancing diversity in generated music through techniques like top-k sampling, nucleus sampling, and temperature parameter control.

The "Motif Transformer" proposed by (Wang *et al.*, 2023) stands out for its focus on generating music with distinct motifs, essential recurring musical patterns. By leveraging a multi-encoder model, which includes an original encoder, a bidirectional long short-term memory-attention encoder, and a gated decoder, this approach excels in capturing motifs, leading to the creation of music with more repetitive fragments. Similarly, the "Theme Transformer" by (Shih *et al.*, 2022) introduces a theme-based conditioning approach for automatic music generation using Transformer models. This approach outperforms baseline models in generating polyphonic pop piano music with repetitions and variations, demonstrating the potential for thematic development and user-generated themes for conditioning. Building upon these advancements, the "Music Transformer" introduced by (Music Transformer: Generating Music with Long-Term Structure, 2018) utilizes a modified relative attention mechanism to generate music compositions with compelling long-term structures, emphasizing the importance of repetition and relative timing. This modification enhances the transformer's ability to capture periodicity in various time scales, paving the way for applications in text processing and audio analysis. However, it is key to note that this is not something that is being sought in a transformer model, as the other side of the coin is looked at to reduce mundane repetition in favour of creating more captivating, musically enriching compositions. In contrast, "Musenet" by (Department of Computer Science, SRM Institute of Science and Technology, Chennai, India. et al., 2020)explores the generation of non-repetitive and enjoyable music using a combination of Support Vector Machines and Neural Nets for discriminatory selection and Generative Pretrained Transformers (GPT2) and LSTMs for music generation. This approach closely mimics human music composition, offering copyright-free music generation based on specific parameters.

The exploration of sampling techniques and temperature parameters in music generation using transformer models has garnered significant attention in recent research. Papers such as (Bjare, Lattner and Widmer, 2023), (Holtzman *et al.*, 2020), (Han *et al.*, 2022) and (Huang *et al.*, 2023) delve into the effectiveness of top-k, nucleus sampling, and temperature sampling to address issues of repetition and diversity in generated music. These studies highlight nucleus sampling as particularly promising for generating high-quality music sequences due to its ability to increase self-similarity and tonal consistency while addressing issues of repetition. Similarly, temperature sampling and top-k sampling are introduced to enrich diversity in music generation, proposing temperature-controlled stochastic sampling methods to balance consistency and diversity. On the other hand, investigations into the temperature parameter'srole in music generation, as highlighted in (Atassi, 2023), (Nuttall, Haki and Jorda, 2021), (Mittal *et al.*, 2021), (Guo, Kang and Herremans, 2023) and (Min *et al.*, 2022) offer valuable insights into controlling the sampling process to balance between avoiding repetition and fostering improvisation. By examining the impact of different temperature settings on the musical output's coherence and diversity, these studies contribute to fine-tuning transformer models for more sophisticated music generation systems. Additionally, research such as (Zhang *et al.*, 2021) and (Zhang *et al.*, 2020) extend these findings to text generation, providing insights into strategies for enhancing diversity in language models through nucleus sampling, top-k sampling, and temperature sampling. These collective efforts mark significant progress in understanding and refining sampling techniques and temperature parameters in transformer models for both music and text generation applications, offering valuable guidance for future developments in the field.

## 2.4 Conclusion

The exploration of artificial intelligence in music production has yielded rich insights, spanning both recurrent neural networks (RNNs) and attention-based transformer models. RNNs, particularly LSTM networks, have played a foundational role in modeling temporal dependencies and sequences, as evidenced by their multifaceted applications in music generation and transcription. Within music generation and transcription, RNNs, notably LSTM networks, play a foundational role, as evidenced by studies like those by (Sigtia, Benetos and Dixon, 2016) and (Goel, Vohra and Sahoo, 2014), which emphasize their importance in modeling sequences and temporal dependencies. Innovations like MusicVAE (Roberts et al., 2019) and novel methodologies for polyphonic music (Boulanger-Lewandowski, Bengio and Vincent, no date) further advance the field. Additionally, genre-specific applications in classical and traditional Celtic music enrich the understanding of RNN versatility. The exploration of novel training techniques and fine-tuning methodologies, as seen in (Jaques et al., 2017), demonstrates RNNs' adaptability in generating high-quality sequences across diverse domains. Overall, these research endeavours underscore the foundational significance of RNNs in capturing temporal dependencies within music and highlight their diverse applications across music transcription, composition, and genre-specific generation, contributing significantly to the evolving landscape of AI-driven music research.

However, recent advancements have seen attention-based transformer models emerge as compelling alternatives to RNNs for capturing long-term dependencies in music generation. Transformer models, initially developed for natural language processing tasks, leverage self-attention mechanisms to process input sequences in parallel, enabling faster computation and improved adaptability to varying input lengths. The landscape of language transformer models illustrates their versatility, with research exploring attention mechanisms, model adaptations, and controllable text generation. Challenges in music generation with transformers, such as computational complexity and fixed-length context limitations, have led to innovations like GPT-2, addressing longer-term dependencies without disrupting temporal coherence. Advancements in multi-sequence music generation introduce models like Calliope and explore conditional multi-track music generation, expanding the scope of transformer-based music modeling. Notable models like MuseNet and Pop Music Transformer demonstrate the transformer's potential to generating diverse and expressive musical compositions. However, challenges remain, including limited human-machine interaction and the need for style representation and emotional expression in music generation. Transformer-GANs emerge as a promising solution, combining transformer architecture with generative adversarial networks to generate music with specific sentiments, albeit with computational complexity and training time drawbacks. Overall, the literature review showcases the evolving landscape of attention-based transformer models in music generation and identifies avenues for future research and development.

Finally, recent advancements in music generation driven by deep learning models, particularly Transformer architectures, have shown promising results in addressing issues such as repetitive patterns and enhancing diversity in generated music. Studies like the "Motif Transformer" (Wang et al., 2023) and the "Theme Transformer" (Shih et al., 2022) demonstrate innovative approaches to capturing distinct motifs and thematic development, respectively. Additionally, research on sampling techniques and temperature parameters, as explored in (Bjare, Lattner and Widmer, 2023), (Holtzman et al., 2020), and others, highlights nucleus sampling as a promising method for generating high-quality music sequences while balancing diversity. Furthermore, investigations into the role of temperature parameters contribute to fine-tuning transformer models for more sophisticated music generation systems. Overall, these collective efforts represent significant progress in understanding and refining transformer-based approaches for music generation, offering valuable insights for future developments in the field.

# Chapter 3: Research Design and Methodology

The aim of the project is to investigate the role of attention-based transformer models in polyphonic music generation, focusing on the influence of training data size and diversity on model performance, the effectiveness of self-attention mechanisms in capturing long-term dependencies, and the model's ability to generate diverse and original music while avoiding repetitive patterns.

## 3.1 Dataset

### 3.1.1 Primary Dataset

In the study conducted on the utilization of attention-based transformer models for polyphonic music generation, four industry experts were interviewed as the primary research participants. An industry expert has been defined as someone who has been working in the music generation industry for at least one year. Industry experts have been 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 has been offered, and a unique perspective on the generated music has been provided, which might not have been noticed otherwise.

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

Due to time constraints, the research has 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 has been chosen for its efficiency in gathering feedback within a short timeframe. Judgement sampling has been selected for exploratory research, allowing for the hand-picking of individuals with relevant experience and proficiency to provide valuable insight into generating music. Experts have been chosen to ensure diversity in expertise, perspectives, and backgrounds, encompassing various job titles across different companies. The research has remained relevant to the overall research objectives, as the experts have offered opinions on the coherence of the music, the impact of the dataset on its quality, and its potential for the music to be monotonous. Their finely tuned ears have detected subtle nuances and intricacies in the music, providing insightful and knowledgeable input while optimizing time and resource utilization.

The method employed in the primary research has involved conducting five expert interviews. Experts have been contacted through LinkedIn and have been invited to participate in the study. The topic of interest, research objectives, and the music generated from the analysis have been 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 has been obtained through email or LinkedIn messages before the interviews have occurred.

The expert interviews have been conducted using a qualitative approach. Depending on logistical concerns, the interviews have been organized for a mutually convenient time and have been performed via Zoom. The experts have 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 have been designed to elicit detailed and insightful responses from the experts, facilitating exploration and clarification of their perspectives.

The audio recordings of the interviews have been transcribed verbatim, and the transcripts have 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 have been employed. The reliability and validity of the results have been ensured through the application of rigorous analysis procedures. Once the analysis has been finished, it has been compared to the findings from secondary research to determine whether the results have aligned with what the experts have found and, if not, to understand the reasons for any disparities.

The preference for the chosen method of primary research has 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 has been essential. This need has been 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 has evaluated whether the newly generated music constitutes a coherent composition that seamlessly integrate with the trained music, distinguishing between a well-structured piece and a mere sequence of random notes. An expert's perspective has been crucial in identifying the strengths and weaknesses of the music, especially regarding the second objective, which has focused on potential quality changes resulting from the use of a new dataset with the model. The expert's keen sense of what constitutes high-quality music has 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 has offered flexibility in shaping the questions posed. If the interviewee has raised intriguing points not previously considered, further exploration has been possible. Likewise, when encountering unclear aspects, the interviewee has been able to provide immediate elaboration. This advantage of the research method has proved invaluable when discussing the third objective, which has 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 have allowed for a deeper exploration and assessment of whether this balance has been achieved.

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

### 3.1.2 Secondary Dataset

Understanding the nuanced relationship between training data characteristics and the performance of attention-based transformer models in polyphonic music generation is paramount for advancing the field. Despite the growing interest in these models, a comprehensive understanding of how the size and diversity of training data influence music quality remains elusive. Addressing this gap, the study focuses on investigating the impact of training data size and diversity on the efficacy of attention-based transformer models in generating polyphonic music. To achieve this first objective, two datasets are meticulously selected, and the output generated from each dataset is evaluated to determine the extent to which variations in training data size and diversity affect the ability of attention-based transformer models to generate high-quality polyphonic music. Through rigorous analysis of the output produced by these models using different datasets, the study aims to provide valuable insights into optimizing training data selection and enhancing the performance of attention-based transformer models in music generation tasks.

The first dataset, POP909 (‘music-x-lab/POP909-Dataset’, 2024), (Wang *et al.*, 2020), encompasses 909 Chinese pop songs with piano arrangements meticulously crafted by professional musicians. This dataset offers a diverse representation of compositions spanning six decades from the 1950s to around 2010 and composed by 462 artists, totalling approximately 60 hours of music. Each arrangement is structured into three tracks (MELODY, BRIDGE, and PIANO), enabling the extraction of lead melody, secondary melodies, and accompaniment. Additionally, the MIDI files in the POP909 dataset have contained expressive dynamics based on the original audio, enhancing the learning process of the models.

In contrast, the second dataset, Lakh MIDI(*The Lakh MIDI Dataset v0.1*, no date), (Raffel, no date)has comprised a vast collection of 176,581 songs across 13 genres, including pop/rock, electronics, country, R&B, jazz, Latin, and more. Most songs have multiple tracks, most of which are aligned with the labelled data in the Million Song Dataset, 11,946 MIDI files have genre labels. A tokenized Lakh dataset has been obtained from Hugging Face to eliminate the duplication of exploring, cleaning, and preprocessing two datasets. This dataset has been tokenized similarly to the POP909 dataset, aiming to achieve marginal impact (*juancopi81/mmm\_track\_lmd\_8bars\_nots · Datasets at Hugging Face*, no date).

When employing the POP909 dataset for symbolic music generation, researchers typically simplify the complexity of the music by restricting selections to songs in the key of C major and with a time signature of 4/4 (Wang *et al.*, 2023), (Guo *et al.*, 2023b), (Hsiao *et al.*, 2021). Upon extensive exploration and analysis of the POP909 dataset, it becomes evident that exclusively focusing on songs in C major with a time signature of 4/4 fails to capture the dataset's inherent diversity. This narrow selection criterion not only drastically reduces the dataset's volume but also neglects the diverse range of musical expressions present across different keys, time signatures, and compositional styles. Thus, it is imperative to preserve the dataset's integrity and inclusivity to foster a thorough understanding of polyphonic music generation. Moreover, breaking the music into chunks has been crucial because it has allowed the model to process information more effectively. By dividing the music into manageable segments, the model has been able to better analyse and understand the intricate details of each segment without becoming overwhelmed by the complexity of the entire piece. This approach has enabled more efficient learning and interpretation of musical patterns, ultimately enhancing the model's ability to generate high-quality polyphonic music.

## 3.2 Data Preprocessing

### 3.2.1 MMM Tokenizer

In the domain of music generation through deep learning models, tokenization plays a pivotal role in transforming raw musical data into a format suitable for computational analysis and modeling. Tokenization involves breaking down complex musical structures, such as melodies, chords, and rhythms, into discrete units or tokens, each representing a specific musical element. These tokens serve as the building blocks for training machine learning models, enabling them to learn and generate music compositions autonomously. Tokenization becomes imperative in music generation projects due to the inherent complexity and richness of musical data comprising intricate patterns, harmonies, and structures. By abstracting away this complexity through tokenization, researchers and composers facilitate the analysis and generation of music compositions with greater ease and efficiency. Furthermore, tokenization allows for the incorporation of domain-specific knowledge and features into the modeling process, guiding the learning process and enhancing the quality of generated music. In essence, tokenization serves as a bridge between the abstract world of music and the computational domain of machine learning, enabling the creation of sophisticated models capable of understanding, analysing, and generating music compositions autonomously, thus paving the way for new creative possibilities in the field of computational music composition.

Exploring the significance of tokenization in music generation projects involved examining various tokenizers used by researchers. The discussion covers the different approaches taken by these scholars in their exploration.

Firstly, (Huang and Yang, 2020) introduce Revamped MIDI-derived Events Tokenizer (REMI), a revamped MIDI-derived events tokenizer tailored for music composition tasks. REMI enhances Transformer-based models by introducing novel event types like Bar and Position, providing explicit information about rhythm and meter. This tokenizer integrates human music knowledge and outperforms traditional methods in composing Pop piano music.

Similarly, (von Rütte *et al.*, 2022) present REMI+, an extension of REMI specifically designed for multi-track modeling. REMI+ enriches the representation with note details, chords, and tempo events, thereby enhancing music modeling. This tokenizer demonstrates proficiency in multi-label feature prediction, offering promising prospects for music generation and analysis.

Moreover,(Hsiao *et al.*, 2021) introduce Compound Word Representation (CP), providing fine-grained control over prediction heads and enabling faster music generation due to reduced time steps. CP's efficiency in training linear Transformer decoders within resource constraints makes it a practical choice for music composition tasks.

In addition, (Ens and Pasquier, 2020b) introduced a Multi-Track Music Machine (MMM), which utilizes the Lakh MIDI Dataset (LMD) to derive token sequences for training GPT2 models aimed at music generation tasks. MMM offers advantages in flexibility and control over the generated output compared to other tokenizers. Its ability to handle both track and bar inpainting methods offers versatility in conditioning the model for generation tasks.

Among the explored tokenizers, MMM has stood out as the best choice for music generation projects. MMM has been a simple yet powerful approach to convert MIDI files to pseudo-words. MMM has provided fine-grained control over prediction heads, enabling precise manipulation of music generation processes. This level of control has allowed for diverse and tailored outputs to be generated with ease, catering to various artistic preferences and requirements. Additionally, MMM has required fewer time steps for music generation compared to other tokenizers, enhancing efficiency and speeding up the model training process. Furthermore, MMM's ability to handle both track and bar inpainting methods has offered versatility in conditioning the model for generation tasks, providing flexibility in generating conditioned or unconditioned music sequences. Overall, MMM's combination of flexibility, efficiency, and control has made it the optimal choice for music generation projects, offering researchers and composers the tools needed to create high-quality and expressive musical compositions.

Open-source libraries such as MidiTok (Fradet *et al.*, 2023) or Musicaiz (Hernandez-Olivan and Beltran, 2023), both of which are MMM tokenizers, are available for tokenizing datasets. Musicaiz has been preferred over MidiTok due to its simplicity and user-friendly interface. Musicaiz has offered a more intuitive workflow, making it accessible to users with varying levels of expertise in MIDI data processing. Additionally, for this project, Musicaiz has proven to be particularly advantageous, as its ease of use has streamlined the tokenization process, allowing for a focus on analysis and experimentation rather than wrestling with complex functionalities.

A diagram of a track

Description automatically generated

Figure 1: The MultiTrack and BarFill representations are shown. The <bar> tokens correspond to complete bars, and the <track> tokens correspond to complete tracks.

The numerical numbers in MMM have matched the MIDI notation's pitch for the notes and instruments. For example, in the given diagram, C4 has been represented by the notation NOTE\_ON=60, and an overdriven guitar has been indicated by INST=30. NOTE\_ON and NOTE\_OFF have indicated the start and stop of a note, and TIME\_DELTA has shown the progression of time on the timeline. The notes have been surrounded in tokens <BAR\_START> and <BAR\_END>, which have been then encased in pseudo-words <TRACK\_START> and <TRACK\_END>, and finally arranged inside tokens <PIECE\_START> and <PIECE\_END>.

### 3.2.2 Training the Tokenizer

Commencing the training process for the GPT-2 transformer model, a crucial decision has emerged regarding the choice of tokenizer. After careful consideration, the adoption of the GPT-2 tokenizer has been deemed essential. This choice streamlines the workload by negating the need to delineate specific parameters regarding tokenization algorithms or the incorporation of special tokens. By adopting the GPT-2 tokenizer, alignment seamlessly with the model architecture intended to be employed is ensured, thus ensuring compatibility and coherence throughout the training process. Moreover, this approach has afforded the flexibility to focus solely on adapting the vocabulary to the specific corpus through the training phase rather than expending resources on customizing tokenizer functionalities. This decision has been reached to optimize efficiency and maintain consistency with established practices within the domain of transformer-based models.

## 3.3 Designing the Transformer Model

In the literature review presented in Chapter 2, various transformer-based models are explored for music generation. GPT-2 emerges as a notable contender alongside models like Recurrent Neural Networks (RNNs), Transformer XL, and Transformer-GANs. Leveraging the transformer architecture introduced by (Vaswani *et al.*, 2017), GPT-2 showcases its capability in natural language processing (NLP) and its adeptness in handling sequential data tasks. Notably, GPT-2 functions as a decoder model in unsupervised learning (*Decoder models - Hugging Face NLP Course*, no date) (Radford *et al.*, no date), GPT-2 excels in processing and comprehending sequential data without explicit task supervision. Its self-attention mechanism, integral to transformer architectures, enables simultaneous analysis of entire input sequences, a departure from the linear processing of traditional models like RNNs.

RNNs, for instance, excel in processing sequential data, making them suitable for musical composition tasks due to their ability to capture temporal dependencies. However, RNNs often encounter issues such as vanishing gradients and struggles in capturing long-term dependencies, which hinder their effectiveness in generating coherent and intricate musical compositions. (Goel, Vohra and Sahoo, 2014), (Eck and Schmidhuber, 2002), (Lalapura, Amudha and Satheesh, 2021) and (Jaques *et al.*, no date) Transformer XL extends the capabilities of traditional RNNs by introducing recurrence mechanisms, addressing challenges like the vanishing gradient problem. Nonetheless, it still grapples with capturing nuanced contextual information essential for music generation tasks, as acknowledged by interviewed experts. (Wu, Wang and Lei, 2020), (Dai *et al.*, 2019) and (Donahue *et al.*, 2019) Transformer GANs represent another avenue, combining the transformer architecture with the generative adversarial network framework to produce diverse and realistic sequences. Despite their promise, Transformer GANs may introduce additional complexity and training challenges compared to standalone transformer models like GPT-2. (Neves, Fornari and Florindo, 2022), (Muhamed *et al.*, 2021b), and (Muhamed *et al.*, 2021a)

GPT-2 Transformer has been selected over alternative options for its unique combination of versatility, generative capabilities, and accessibility. Developed by OpenAI (Radford *et al.*, no date), GPT-2 stands out for its transformer-based architecture, which has not only excelled in natural language processing tasks but has also adapted well to various domains, including music generation. Unlike RNNs, GPT-2 has overcome the limitations associated with capturing long-term dependencies while also offering a more accessible and well-supported framework compared to newer architectures like Transformer XL and Transformer GANs. Moreover, GPT-2's pre-trained weights on a vast corpus of text data have simplified the fine-tuning process for music generation tasks, reducing computational overhead and training time. Its robust generative capabilities have enabled the production of coherent sequences of tokens essential for crafting musical compositions with structure and coherence like those by human composers. Furthermore, GPT-2's availability in popular deep learning frameworks such as TensorFlow and PyTorch has facilitated seamless integration into the research workflow, enabling swift experimentation and iteration. (Banar and Colton, 2022), (Banar and Colton, 2021), (Imasato *et al.*, 2023), and (Guo *et al.*, 2023a)

### 3.3.1 Decoder Transformer Architecture

GPT-2 operates as a decoder model within the transformer architecture framework. Unlike the original transformer design, which incorporates both encoder and decoder components (Vaswani *et al.*, 2017), GPT-2 exclusively employs the decoder segment for its tasks. This decoder architecture enables GPT-2 to generate outputs sequentially, relying solely on its previously generated tokens without access to any specific input sequence during the generation process.

During inference, GPT-2 initiates with an initial token, typically a special token marking the beginning of a sequence. Subsequently, it generates subsequent tokens autoregressively, one at a time, guided by the probabilities assigned to each token by the model. At each step, the model attends to its own previously generated tokens to inform the generation of the next token in the sequence. This autoregressive decoding approach enables GPT-2 to craft contextually relevant and coherent outputs that maintain the overall structure and style of the input data it was trained on.

Primarily functioning as a decoder model, GPT-2 excels in tasks like language modeling, text generation, and other sequence generation tasks such as music generation. Leveraging the transformer architecture's capability to process and generate sequential data, GPT-2's ability to generate high-quality outputs based on learned context has solidified its prominence in natural language processing and beyond (*Decoder models - Hugging Face NLP Course*, no date).

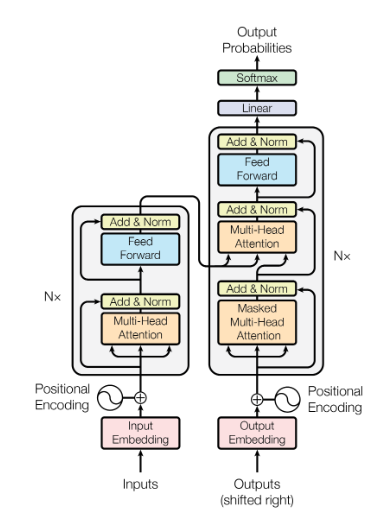


Figure 2: Original Transformer model architecture from Attention Is All You Need. The encoder is depicted on the right-hand side and the decoder on the left-hand side (Vaswani et al., 2017)

## 3.4 Self-attention

The ability of GPT-2 to process sequences in parallel enables it to capture dependencies between tokens, regardless of their sequence positions, thereby overcoming challenges related to long-range dependencies. In contrast, traditional RNNs frequently encounter difficulties in capturing such dependencies because of the vanishing gradient problem. This issue hinders the effective propagation of information from distant tokens throughout the network during training. Consequently, RNNs may struggle to maintain contextual coherence over long sequences. In contrast, GPT-2's self-attention mechanism allows it to overcome this limitation by explicitly modeling the relationships between all tokens in the input sequence, regardless of their distance from each other. This enables the model to effectively handle long-range dependencies and generate more coherent and contextually relevant outputs.

### 3.4.1 Multi-Head Attention

In GPT-2, the utilization of multi-head attention is a key component of its transformer architecture, contributing significantly to its ability to process and understand sequential data effectively. Multi-head attention involves applying the attention mechanism multiple times in parallel, each with its own set of weight matrices. This mechanism enables the model to focus on different parts of the input sequence simultaneously, thereby enhancing its capacity to capture diverse patterns and relationships within the data.

By employing multiple heads of attention, GPT-2 can attend to different aspects or subspaces of the input sequence concurrently. Each head learns to attend to different parts of the sequence and extract relevant information independently, enabling the model to capture a broader range of contextual information. This parallel processing of attention heads allows GPT-2 to effectively encode complex relationships and dependencies within the input sequence, leading to more robust representations of the data.

Additionally, the use of multiple heads of attention provides the model with redundancy and robustness against overfitting. Each head learns different representations of the input data, which are then aggregated and combined to form a comprehensive representation. This ensemble of attention heads ensures that the model can capture diverse patterns and nuances present in the data, leading to more robust and generalizable predictions. (Vaswani *et al.*, 2017)

To accomplish the second objective of evaluating the impact of the self-attention mechanism on the music quality, a number of approaches were considered.

One approach involves comparing the performance of the GPT-2 model with self-attention against baseline models such as RNN. Controlled experiments and comparative analyses are conducted to assess whether self-attention provides benefits in terms of model performance metrics such as accuracy, perplexity, or generation quality. While this comprehensive assessment enables a thorough evaluation of the contribution of self-attention to the overall success of the transformer architecture, specifically in the context of music generation tasks, it may require significant computational resources and time for training and analysing for both models.

Another strategy entails fine-tuning hyperparameters related to self-attention, such as dropout rates, learning rates, and warm-up ratios. This process provides further insights into its effectiveness. However, hyperparameter tuning experiments can be computationally intensive and may require extensive experimentation to identify optimal settings. Additionally, the search space for hyperparameters can be vast, leading to challenges in finding the most suitable configurations.

Adjusting the number of attention heads in the model architecture offers another avenue for evaluating the self-attention mechanism. By varying these parameters, observers can observe how changes in the structure of the model affect its ability to capture dependencies within input sequences. Training models with different numbers of attention heads allow for a systematic exploration of the impact of self-attention on learning dynamics and model behaviour.

The decision has been made to adjust the number of attention heads in the model architecture to evaluate the self-attention mechanism due to several reasons. Firstly, it has allowed for a more granular analysis of how the architecture affects the model's ability to capture dependencies within input sequences. By systematically varying the number of attention heads, observers have noted how changes in this aspect of the model structure influence its performance in terms of capturing long-term dependencies. Additionally, this approach has provided insights into the scalability of the self-attention mechanism; understanding how the number of attention heads has impacted the model's behaviour has informed decisions regarding model complexity and computational efficiency. Moreover, adjusting attention heads has been less computationally intensive compared to other methods, such as fine-tuning hyperparameters or comparing against baseline models. Thus, it has offered a practical and effective way to evaluate the self-attention mechanism's effectiveness in capturing long-term dependencies, contributing to a deeper understanding of transformer model behaviour in the context of music generation tasks.

## 3.5 Repetitive patterns

To address the third objective of evaluating the diversity and originality of polyphonic music generated by attention-based transformer models, various options were explored, including sampling techniques and temperature parameters. The literature review in Chapter 3 revealed several prominent approaches used in recent research. Studies by (Bjare, Lattner and Widmer, 2023), (Holtzman *et al.*, 2020), (Han *et al.*, 2022), and (Huang *et al.*, 2023) have extensively examined top-k, nucleus sampling, and temperature sampling to mitigate issues of repetition and enhance diversity in generated music.

Nucleus sampling emerged as a promising method for generating high-quality music sequences by increasing self-similarity and tonal consistency while effectively addressing repetition issues. Additionally, temperature sampling and top-k sampling have been introduced to enrich diversity in music generation, offering temperature-controlled stochastic sampling methods to balance consistency and diversity.

Furthermore, investigations into the role of temperature parameters in music generation, as highlighted by (Atassi, 2023), (Nuttall, Haki and Jorda, 2021), (Mittal *et al.*, 2021), (Guo, Kang and Herremans, 2023), and (Min *et al.*, 2022), have provided valuable insights into controlling the sampling process. These studies have explored the impact of different temperature settings on the coherence and diversity of musical output, contributing to the fine-tuning of transformer models for more sophisticated music generation systems.

Moreover, research by (Zhang *et al.*, 2021) and (Zhang *et al.*, 2020) has extended these findings to text generation, offering insights into strategies for enhancing diversity in language models through nucleus sampling, top-k sampling, and temperature sampling.

Considering the breadth and depth of research findings and experimentation of all three approaches, the focus has been placed on temperature parameters as the primary option for enhancing the diversity and originality of polyphonic music generated by attention-based transformer models. By varying the temperature parameter, adjustments could be made to the level of randomness in the sampling process, thereby influencing the diversity and novelty of the generated music. This exploration has enabled the observation of how different temperature settings achieve a balance between avoiding repetition and fostering improvisation. Through examining the impact of various temperature settings on musical output coherence and diversity, efforts have been made to refine transformer models for more nuanced and creative music generation, thereby contributing to the advancement of machine learning models in this domain.

## 3.6 Hyperparameters Optimization

The Custom GPT2 model has been trained on a 7-GPU NVIDIA A100 Tensor Core GPU system with the following hyperparameters listed in Table 1.

|  |  |
| --- | --- |
| **Hyperparameter** | **Value** |
| Number of layers (n\_layer) | 6 |
| Number of attention heads (n\_head) | 8 |
| Embedding dimension (n\_emb) | 512 |
| Number of training epochs | 1 |
| Training batch size per device | 8 |
| Evaluation batch size per device | 4 |
| Evaluation strategy | 350 |
| Checkpoint save strategy | 350 |
| Evaluation steps | 350 |
| Logging steps | 350 |
| Log the first step | TRUE |
| Save total limit | 5 |
| Save steps | 350 |
| Learning rate scheduler type | Cosine |
| Initial learning rate | 5.00E-04 |
| Warmup ratio | 0.01 |
| Weight decay coefficient | 0.01 |
| Random seed for reproducibility | 1 |
| Load best model at end of training | TRUE |

Table 1: The hyperparameter values used in the GPT-2 model training

These hyperparameters were chosen based on experimentation and domain knowledge to achieve optimal model performance while considering computational resources and training time.

## 3.7 Model Evaluation Metrics

The GPT-2 model has been evaluated using training loss and validation loss metrics over multiple training steps to track its progress. These metrics have served as indicators of the model's performance during the training and evaluation phases, respectively. The training loss has measured how well the model has fit the training data, indicating how effectively it has learned from the provided dataset. A decreasing training loss has suggested that the model has been improving its ability to predict the target outputs. On the other hand, the validation loss has assessed the model's performance on unseen data. This metric has helped prevent overfitting by monitoring the model's ability to generalize to new data. A validation loss comparable to the training loss has indicated that the model has learned useful patterns from the training data without memorizing it, suggesting good generalization performance. The decreasing trend in both training and validation losses over time has suggested that the model has been effectively learning from the data and improving its performance.

# Chapter 4: Results and Discussion

The listening test has been selected as a method to evaluate the AI music generated from the GPT-2 model with the objectives in mind. Firstly, it aligns with the objective of assessing the ability of attention-based transformer models to produce diverse and original polyphonic music, as outlined in the third objective. By directly involving human participants in listening to and evaluating the music, qualitative feedback on factors such as diversity, originality, and the avoidance of repetitive patterns has been gathered, which are crucial aspects of musical quality and creativity. Additionally, the listening test has tackled the second objective by enabling the evaluation of the effectiveness of self-attention mechanisms in capturing long-term dependencies in the music. Participants' perceptions of coherence and the overall structure of the compositions have provided insights into how well the model has captured these long-term relationships between musical elements. Moreover, considering the first objective, which has focused on the impact of training data on model performance, the listening test has allowed the evaluation of the quality of generated music across different datasets, thereby offering valuable insights into the relationship between training data size and diversity and the musical quality of the generated compositions. Overall, the listening test has served as a comprehensive and human-centred approach to evaluating the AI-generated music from the GPT-2 model, addressing multiple objectives and providing valuable insights into its musical creativity and performance. Furthermore, the use of listening tests has been a crucial part of evaluating music in several research papers, including (Imasato *et al.*, 2023), (Ji, Luo and Yang, 2020), (Choi *et al.*, 2021), (Makris, Agres and Herremans, 2021), (Zhao *et al.*, 2022), (Hsiao *et al.*, 2021), (Hung *et al.*, 2021), (Ferreira, Limongi and Fávero, 2023), (Jin *et al.*, 2022), (Huang and Yang, 2020), and (Shih *et al.*, 2022). This underscores the importance of such tests in comprehensively evaluating AI-generated music across various contexts and methodologies.

The listening test was developed and distributed through Google Forms. For the listening test, two types of populations were initially selected: people who have a musical background (e.g., play an instrument, studied music, etc.) and people who do not have a musical background. The criterion adopted for choosing the populations was based on the hypothesis that people with a musical background have a greater ability to distinguish the subtleties in the differences in the music samples. The purpose of this criterion was to add details for the final analysis of the research. All subjects, totalling 33 participants (23 with no musical background and 10 with a musical background), had a total of thirteen musical samples to be heard, where two of them derived from a composition written by a human. A musical sample was collected from the model output after each change was made to the model. The musical samples were the same for the entire tested population. Additionally, participants were asked if they were 18 years of age or older and if they had any medical constraints such as hearing loss or hearing sensitivity. Only participants who were over 18 and without medical constraints were included in our results. Moreover, participants were asked to specify the type of headphones they were wearing (In-ear, On-ear, Over-ear) to see if this impacted the results.

The following questions, presented for each musical sample in Section 2, have addressed the third objective of evaluating the diversity and originality of polyphonic music generated by attention-based transformer models. To facilitate this evaluation, the temperature parameter has been adjusted, and the resulting music samples have been saved using the weights and bias experiment tracker then attached to the Google Form. The adjustment of the temperature parameter to values of 0.25, 0.5, 0.75, and 1 has aimed to examine how different levels of randomness have affected the diversity and creativity of generated music. This variation in model output conditions has allowed for a comprehensive exploration of the model's capabilities and performance across a range of settings, aligning with the research objective to assess its ability to avoid repetitive patterns and generate novel musical ideas.

1. How would you rate the originality of the musical ideas presented in the sample generated by the computer model? (Low, Moderate, High)

Figure 3: Musical Participants Evaluating Originality of Music Sample

Figure 4: Non-Musical Participants Evaluating Originality of Music Sample

As evidenced by the blue segment in the above two stacked area graphs, participants with a musical background have not identified a significant improvement in the originality of the music samples as the temperature parameter has been increased. In contrast, non-musical participants have seen a steady increase. This observation may be attributed to the larger number of non-musical participants, thus suggesting that the observed trend is influenced by the composition of the participant groups.

1. Did you detect any repetitive patterns in the music samples generated by the computer model? ("repetitive patterns in music" refers to uninteresting and predictable sequences that lack creativity or variety, resulting in a diminished listening experience) (Yes/No)

Figure 5: Musical Participants Detecting Repetitive Patterns

Figure 6: Non-Musical Participants Detecting Repetitive Patterns

As observed in the above stacked bar graph, the music participants have shown a lack of sensitivity to reductions in repetitive patterns in the music samples, contrasting with the response of non-musical participants. However, both groups have concurred that repetitive patterns have diminished as the temperature parameter has increased. This validates the hypothesis that attention-based transformer models have generated diverse and original polyphonic music, specifically reduced repetitive patterns and introduced novel musical ideas.

1. Were there instances where the computer model successfully avoided repetitive patterns and introduced unique musical ideas? ("repetitive patterns in music” refers to uninteresting and predictable sequences that lack creativity or variety, resulting in a diminished listening experience) (Yes/No)

Figure 7: Musical Participants Evaluating the Uniqueness of Music Ideas

Figure 8: Non-Musical Participants Evaluating Uniqueness of Music Ideas

As evident in the above stacked column charts, musical participants have perceived that the music has avoided repetitive patterns and introduced unique musical ideas more prominently as the temperature parameter increased. Similarly, non-musical participants have also shown this trend, although they were less discerning of the lower temperature parameter's tendency towards repetition compared to musical participants. However, notably, both groups have recognized that the temperature parameter at 0.75 has been most effective in avoiding repetitive patterns and introducing unique musical ideas. This observation aligns with findings from other research papers.

1. How musically engaging do you find the samples generated by the computer model? (Not engaging, Moderately engaging, Very engaging)

Figure 9: Musical Participants Evaluating Perceived Musical Engagement Levels

Figure 10: Non-Musical Participants Evaluating Perceived Musical Engagement Levels

In the stack column charts above, participants with a musical background find the music sample with a temperature parameter of 0.75 the most engaging, as expected. Sample 1 and 4 with temperature parameters of 0.25 and 1 ,respectively, are also identified as the second most engaging. This is expected for sample 4 but not for sample 1. Sample 2 is only moderately engaging, confirming expectations. Conversely, participants with a non-musical background notice that the musical samples become progressively more engaging as the temperature parameter increases, which aligns with expectations.

The following questions have been proposed for each musical sample in Section 3, which has covered the second objective of evaluating the effectiveness of self-attention mechanisms in transformer models for capturing long-term dependencies in polyphonic music. To facilitate this evaluation, the number of attention heads has been adjusted, and again, resulting music samples have been saved using the weights and bias experiment tracker and attached to the Google Form. The adjustment of the number of attention heads to values of 1, 2, 4, 8, and 16 has allowed for a thorough investigation into how varying levels of attention have affected the model's ability to capture long-term dependencies in the music.

1. Did you hear any parts where the music sounded organized, like a well-put-together song, rather than just random notes? (Yes/No)

Figure 11: Music Participants Evaluating the Coherence of Music Sample

Figure 12: Non-Music Participants Evaluating the Coherence of Music Sample

When examining the orange section in the above stacked area charts, participants with a musical background feel that sample 6 with 2 attention heads sounds the most organized and well put together. On the other hand, those without a music background consider sample 8 with 8 attention heads to be more organized and well put together, followed by sample 9 with 16 attention heads. Once more, participants without a musical background align more closely with expectations. With more musical participants, this difference is more likely to diminish.

1. How musically pleasing did you find the music samples generated by the computer model? (Not pleasing, Moderately pleasing, Very pleasing)

Figure 13: Musical Participants Rating of Pleasure in Music Samples

Figure 14: Non-Musical Participants Rating of Pleasure in Music Samples

When focusing on the blue bars in the clustered column graphs above, it becomes evident that as the attention heads increase, so does the rating of how musically pleasing the music samples are. This holds true for both participants with and without a musical background. This observation confirms the hypothesis that self-attention mechanisms in transformer models effectively capture long-term dependencies in polyphonic music, contributing to the overall perceived musical satisfaction regardless of participants' musical backgrounds.

The following questions have been proposed for each musical sample in Section 4, which covers the second objective of assessing 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. To facilitate this evaluation, a music sample generated by the POP909 data with the temperature parameter of 0.75 and the number of attention heads at 8 has been used to compare against one of the songs from the POP909 dataset. This was likewise done for the Lakh dataset. Again, the resulting music samples that have been generated by the GPT-2 model have been saved using the weights and bias experiment tracker and attached to the Google Form. This process has allowed for an analysis of how the size and diversity of the training data impact the quality of the generated music, addressing the first objective of the study.

1. Did you notice any fluctuations in the quality of the music you just heard? (Yes/No)

Figure 15: Musical Participants Evaluating the Quality of the Music Sample

Figure 16: Non-Musical Participants Evaluating the Quality of the Music Sample

Focusing on the blue segments in the stacked bar charts above, musical participants provide a more decisive evaluation of the quality of the music samples compared to non-musical participants. Musical participants consistently believe that the larger dataset produces higher quality music than the smaller dataset, POP909. This observation confirms the hypothesis that the size of the dataset directly influences the perceived quality of the generated music.

1. How would you rate the overall quality of the music you just heard? i.e. 5 being excellent and 1 being very poor (1,2,3,4,5)

Figure 17: Musical Participants Rating of Overall Music Sample Quality

Figure 18: Non-Musical Participants Rating of Overall Music Sample Quality

When asked about the overall quality of the music samples, the original music samples from both datasets received higher ratings than the music generated by the transformer model. This consensus is confirmed by both populations, indicating that more effort is required to elevate the quality of the generated music to match that of the original datasets. Potential approaches to achieve this include implementing more customized tokenization, utilizing sweep by Weights and Bias for more accurate hyperparameter tuning of the model, or exploring alternative transformer architectures.

1. Did the music you just listened to flow smoothly and make sense to you? (Yes/No)

Figure 19: Musical Participants Perception of Music Flow and Coherence

Figure 20: Non-Musical Participants Perception of Music Flow and Coherence

In the cluster column graphs above, it is evident for both populations that the original music samples are perceived to exhibit musical flow and coherence more effectively than the generated music. Moreover, for both populations, it is apparent that the larger dataset (Lakh) demonstrates greater musical flow and coherence compared to the smaller dataset (POP909). This observation confirms the hypothesis that the size of the dataset significantly impacts the level of musical flow and coherence perceived in the generated music by both populations.

## Chapter 5: Ethical and Legal Considerations

### 5.1 Primary Research

Ethical considerations have been central to the data analysis project. Voluntary participation and informed consent from all five participating experts have been required for the primary research. This has been 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 has been encouraged, with full respect for their decision to participate or withdraw from the study at any stage.

Trust and respect have been of the utmost importance for all participants in the study. This has 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 have been valued.

Ensuring the accuracy of result reporting has been paramount. This process entailed transcribing the recordings verbatim and extracting vital themes, quotes, and interesting insights while avoiding the inclusion of redundant or irrelevant information.

When composing the thesis results, confidentiality and anonymity of the experts' responses have been 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., has been 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 have been presented in a manner that honours the experts' contributions and upholds the research's integrity. Themes and insights derived from all five interviews have been summarized, enhancing readability and preserving the essence of each interviewee's input.

Overall, ethical principles have taken precedence in the thesis project. Voluntary participation and informed consent have been 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 have played a central role, emphasized through punctuality, preparedness, and genuine appreciation for the experts' opinions and expertise. Reporting and dissemination have followed 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 have received aliases and personal information has been securely stored. Ultimately, the findings have been 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 have ensured a robust and respectful execution of the primary research in the thesis project.

### 5.2 Secondary Research

In the ever-evolving landscape of AI-generated music, a fundamental consideration has lain in harmonious coexistence with established copyrights and the avoidance of infringement. Irish, UK, and USA copyright laws have 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 has liberated the music, letting it be used, changed, and shared without worrying about breaking copyright laws. As the exploration of AI-driven musical composition has continued, it has become pivotal to not only respect these legal frameworks but also to leverage open-source data discussed in the Data Collection section. This data has aligned with copyright regulations and has duly acknowledged the sources, thereby safeguarding intellectual property and artistic integrity.

When venturing into the domain of AI-generated music, it has been 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 has been far from straightforward. Their development has been influenced by a complex web of factors, often intersecting and shaping the transformation of musical styles and practices. Music has been profoundly shaped by elements such as geography, culture, religion, historical events, technological advancements, and 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 have been the potential presence of stereotypes and cultural biases in the music it has generated. Stereotypes could have emerged when the AI model, intentionally or unintentionally, has replicated simplistic or biased notions about certain musical styles, genres, or cultures. For example, the model might have falsely associated specific musical elements with cultural clichés, resulting in misrepresentations. These stereotypes could have perpetuated cultural insensitivity and have led to feelings of hurt or disrespect among listeners.

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

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

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

In the dynamic realm of AI-generated music, maintaining harmony with copyright laws, avoiding biases, and crafting emotionally engaging compositions have been essential for shaping the future of this art form. In the complex world of musical creation, a firm dedication to legal and ethical standards has remained unwavering, alongside an embrace of the limitless potential that AI provides for creative expression. Upholding the values of respect, diversity, and emotional connection has ensured that AI-generated music has not only honoured the past but also paved the way for a harmonious and inclusive musical future.

# Chapter 6: Conclusion

In this thesis, the complex landscape of AI-generated music has been tackled, focusing on attention-based transformer models for polyphonic music generation. Three primary objectives have been addressed: understanding the impact of training data on model performance, evaluating the effectiveness of self-attention mechanisms in capturing long-term dependencies, and assessing the models' ability to generate diverse and original music while avoiding repetitive patterns.

The exploration of artificial intelligence in music production has yielded significant insights, with attention-based transformer models emerging as promising alternatives to traditional recurrent neural networks (RNNs) for capturing long-term dependencies in music. While RNNs, notably LSTM networks, have historically been crucial for modeling temporal dependencies within music, recent advancements in transformer architectures have demonstrated their potential in addressing issues such as repetitive patterns and enhancing diversity in generated music. Transformer models, initially developed for natural language processing tasks, leverage self-attention mechanisms to process input sequences in parallel, facilitating faster computation and improved adaptability to varying input lengths. These models have showcased versatility in language tasks, leading to research exploring attention mechanisms, model adaptations, and controllable text generation. Challenges in applying transformers to music generation, such as computational complexity and fixed-length context limitations, have prompted innovations like GPT-2, which effectively address longer-term dependencies without sacrificing temporal coherence.

This thesis delves into the intricate dynamics involved in training attention-based transformer models for polyphonic music generation. By meticulously analysing diverse datasets and employing sophisticated tokenization techniques, the study sheds light on the impact of training data characteristics on model performance. The chosen GPT-2 transformer model, with its self-attention mechanism and decoder architecture, proves effective in generating quality polyphonic music. The thesis highlights the significance of self-attention and multi-head attention mechanisms in capturing long-term dependencies and enhancing music diversity and originality. Moreover, the exploration of sampling techniques offers nuanced control over the creativity of model-generated music, contributing to the refinement of transformer models for music generation tasks. The human-centred approach, employing listening tests, provides qualitative feedback on music quality, diversity, and coherence, validating the efficacy of the proposed methodologies.

Through meticulous analysis and experimentation, valuable insights into these objectives have been uncovered. It has been found that the size and diversity of training data significantly influence the quality of generated music, highlighting the importance of comprehensive datasets in achieving higher-quality outputs. Moreover, the evaluation of self-attention mechanisms has revealed their effectiveness in capturing long-term dependencies, essential for producing coherent and musically pleasing compositions. Additionally, attention-based transformer models have been shown to excel in generating diverse and original music while mitigating repetitive patterns, showcasing their potential for fostering creativity and innovation in algorithmic music composition. The significance of this work extends beyond technical advancements. By employing a human-centred approach through listening tests, qualitative feedback on music quality and perception has been gained, underscoring the subjective nature of musical evaluation. Furthermore, ethical considerations have ensured the integrity and respectfulness of the research, emphasizing voluntary participation, informed consent, and confidentiality.

Overall, this thesis makes significant strides in advancing the field of AI-generated music by providing valuable insights into the capabilities and limitations of attention-based transformer models. Through a focus on fostering creativity, innovation, and ethical practices, the research contributes to shaping a future in AI-generated music that prioritizes legal, cultural, and ethical considerations while promoting musical expression.

### 6.1 Future Research

Considering the findings and limitations observed in this study, several avenues for future research warrant exploration. Firstly, broadening the demographic scope of participants could yield deeper insights into how individuals' musical backgrounds influence their evaluation of music samples. Understanding whether participants' musical preferences impact their perception of the samples would provide valuable context. Moreover, delving into alternative transformer architectures and experimenting with diverse hyperparameter settings could potentially bolster the performance of attention-based models in generating varied and original polyphonic music. These explorations could lead to significant advancements in computer-generated music, fostering greater creativity and expressiveness in algorithmic composition. Additionally, considering the possibility of survey fatigue among participants due to the presentation of 34 music samples, despite their brevity, is crucial. Implementing tester questions to gauge participant attentiveness throughout the survey process could offer insights and mitigate potential fatigue-induced biases in responses. Furthermore, integrating a quantitative evaluation of the music samples would provide a more comprehensive assessment of their quality. Comparing the original music with the generated compositions using established metrics would offer nuanced insights into the model's performance and facilitate iterative improvements prior to survey distribution. Exploring the influence of different headphone types on music perception, particularly with a larger participant pool, could offer valuable insights into how hardware affects participants' experiences. Additionally, ensuring uniform survey conditions, such as standardizing the environment and equipment, would enhance result reliability and minimize external factors' influence. Addressing these considerations could significantly enhance future research endeavours, advancing comprehension of computer-generated music and refining evaluation methodologies employed in the field.

# References

Atassi, L. (2023) ‘Musical Form Generation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2310.19842.

Banar, B. and Colton, S. (2021) ‘Generating Music with Extreme Passages using GPT-2’.

Banar, B. and Colton, S. (2022) ‘A Systematic Evaluation of GPT-2-Based Music Generation’, in, pp. 19–35. Available at: https://doi.org/10.1007/978-3-031-03789-4\_2.

Bjare, M.R., Lattner, S. and Widmer, G. (2023) ‘Exploring Sampling Techniques for Generating Melodies with a Transformer Language Model’. arXiv. Available at: https://doi.org/10.48550/arXiv.2308.09454.

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.

Choi, K. *et al.* (2021) ‘Chord Conditioned Melody Generation With Transformer Based Decoders’, *IEEE Access*, 9, pp. 42071–42080. Available at: https://doi.org/10.1109/ACCESS.2021.3065831.

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.

*Decoder models - Hugging Face NLP Course* (no date). Available at: https://huggingface.co/learn/nlp-course/chapter1/6 (Accessed: 19 February 2024).

Department of Computer Science, SRM Institute of Science and Technology, Chennai, India. *et al.* (2020) ‘Musenet : Music Generation using Abstractive and Generative Methods’, *International Journal of Innovative Technology and Exploring Engineering*, 9(6), pp. 784–788. Available at: https://doi.org/10.35940/ijitee.F3580.049620.

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.

Eck, D. and Schmidhuber, J. (2002) ‘Finding temporal structure in music: blues improvisation with LSTM recurrent networks’, *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing* [Preprint]. Available at: https://doi.org/10.1109/NNSP.2002.1030094.

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

Ens, J. and Pasquier, P. (2020b) ‘MMM : Exploring Conditional Multi-Track Music Generation with the Transformer’. arXiv. Available at: http://arxiv.org/abs/2008.06048 (Accessed: 14 February 2024).

Ferreira, P., Limongi, R. and Fávero, L.P. (2023) ‘Generating Music with Data: Application of Deep Learning Models for Symbolic Music Composition’, *Applied Sciences*, 13(7), p. 4543. Available at: https://doi.org/10.3390/app13074543.

Fradet, N. *et al.* (2023) ‘miditok: A Python package for MIDI file tokenization’. arXiv. Available at: http://arxiv.org/abs/2310.17202 (Accessed: 17 February 2024).

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.* (2023a) ‘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.

Guo, Y. *et al.* (2023b) ‘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.

Guo, Z., Kang, J. and Herremans, D. (2023) ‘A Domain-Knowledge-Inspired Music Embedding Space and a Novel Attention Mechanism for Symbolic Music Modeling’, *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4), pp. 5070–5077. Available at: https://doi.org/10.1609/aaai.v37i4.25635.

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.

Han, S. *et al.* (2022) ‘Symbolic Music Loop Generation with Neural Discrete Representations’. arXiv. Available at: https://doi.org/10.48550/arXiv.2208.05605.

Hernandez-Olivan, C. and Beltran, J.R. (2023) ‘Musicaiz: A python library for symbolic music generation, analysis and visualization’, *SoftwareX*, 22, p. 101365. Available at: https://doi.org/10.1016/j.softx.2023.101365.

Holtzman, A. *et al.* (2020) ‘The Curious Case of Neural Text Degeneration’. arXiv. Available at: http://arxiv.org/abs/1904.09751 (Accessed: 22 August 2023).

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, W. *et al.* (2023) ‘Hyperbolic Music Transformer for Structured Music Generation’, *IEEE Access*, 11, pp. 26893–26905. Available at: https://doi.org/10.1109/ACCESS.2023.3257381.

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).

Hung, H.-T. *et al.* (2021) ‘EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2108.01374.

Imasato, N. *et al.* (2023) ‘Using a Language Model to Generate Music in its Symbolic Domain while Controlling its Perceived Emotion’, *IEEE Access*, pp. 1–1. Available at: https://doi.org/10.1109/ACCESS.2023.3280603.

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’.

Ji, S., Luo, J. and Yang, X. (2020) ‘A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions’. arXiv. Available at: http://arxiv.org/abs/2011.06801 (Accessed: 18 November 2023).

Jin, C. *et al.* (2022) ‘MetaMGC: a music generation framework for concerts in metaverse’, *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1), p. 31. Available at: https://doi.org/10.1186/s13636-022-00261-8.

*juancopi81/mmm\_track\_lmd\_8bars\_nots · Datasets at Hugging Face* (no date). Available at: https://huggingface.co/datasets/juancopi81/mmm\_track\_lmd\_8bars\_nots (Accessed: 23 February 2024).

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.

Lalapura, V.S., Amudha, J. and Satheesh, H.S. (2021) ‘Recurrent Neural Networks for Edge Intelligence: A Survey’, *ACM Computing Surveys*, 54(4), p. 91:1-91:38. Available at: https://doi.org/10.1145/3448974.

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.

Min, J. *et al.* (2022) ‘Music Generation System for Adversarial Training Based on Deep Learning’, *Processes*, 10(12), p. 2515. Available at: https://doi.org/10.3390/pr10122515.

Mittal, G. *et al.* (2021) ‘Symbolic Music Generation with Diffusion Models’. arXiv. Available at: http://arxiv.org/abs/2103.16091 (Accessed: 4 December 2023).

Muhamed, A. *et al.* (2021a) ‘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.

Muhamed, A. *et al.* (2021b) ‘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).

*Music Transformer: Generating Music with Long-Term Structure* (2018) *Magenta*. Available at: https://magenta.tensorflow.org/music-transformer (Accessed: 16 July 2023).

‘music-x-lab/POP909-Dataset’ (2024). Music X Lab. Available at: https://github.com/music-x-lab/POP909-Dataset (Accessed: 16 February 2024).

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’.

Nuttall, T., Haki, B. and Jorda, S. (2021) ‘Transformer Neural Networks for Automated Rhythm Generation’, in *International Conference on New Interfaces for Musical Expression*. *NIME 2021*. Available at: https://doi.org/10.21428/92fbeb44.fe9a0d82.

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).

*Performance RNN: Generating Music with Expressive Timing and Dynamics* (2017) *Magenta*. Available at: https://magenta.tensorflow.org/performance-rnn (Accessed: 16 July 2023).

Radford, A. *et al.* (no date) ‘Language Models are Unsupervised Multitask Learners’.

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.

Raffel, C. (no date) ‘Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching’.

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).

von Rütte, D. *et al.* (2022) ‘FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control’. arXiv. Available at: https://doi.org/10.48550/arXiv.2201.10936.

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).

Shih, Y.-J. *et al.* (2022) ‘Theme Transformer: Symbolic Music Generation with Theme-Conditioned Transformer’. arXiv. Available at: https://doi.org/10.48550/arXiv.2111.04093.

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.

*The Lakh MIDI Dataset v0.1* (no date). Available at: https://colinraffel.com/projects/lmd/#license (Accessed: 21 February 2024).

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, H. *et al.* (2023) ‘Motif Transformer: Generating Music With Motifs’, *IEEE Access*, 11, pp. 63197–63204. Available at: https://doi.org/10.1109/ACCESS.2023.3287271.

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.

Wang, Z. *et al.* (2020) ‘POP909: A Pop-song Dataset for Music Arrangement Generation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2008.07142.

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.

Zhang, H. *et al.* (2020) ‘Trading Off Diversity and Quality in Natural Language Generation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2004.10450.

Zhang, X. *et al.* (2021) ‘Improving Diversity of Neural Text Generation via Inverse Probability Weighting’. arXiv. Available at: http://arxiv.org/abs/2103.07649 (Accessed: 13 February 2024).

Zhao, Z. *et al.* (2022) ‘A Review of Intelligent Music Generation Systems’. arXiv. Available at: http://arxiv.org/abs/2211.09124 (Accessed: 25 May 2023).

# Appendix

## Appendix A: Github Link

https://github.com/sba22222/Thesis-main

## Appendix B: Open-Source Data Permissions

These references have been added to the Research Design and Methodology under the Dataset section.

POP909 Dataset

A close-up of a message

Description automatically generated

Lakh Dataset

A screenshot of a paper

Description automatically generated

## Appendix C: Survey Link

<https://docs.google.com/forms/d/e/1FAIpQLSdQZ9Os7_MzWvTaznZHw-76SVnfuZo4jZWoGvQt83fcoz7WIg/viewform?usp=sf_link>

## Appendix D:Interview Permissions

Participant 1

A screenshot of a email

Description automatically generated

Participant 2

A screenshot of a chat

Description automatically generated

Participant 3

A screenshot of a message

Description automatically generated

Participant 4

A screenshot of a chat

Description automatically generated

## Appendix E: Interview Transcripts

### Participant 1

WhatsApp Audio 2023-08-01 at 4.30.15 PM.mp4

Transcript

Speaker 2

So I don't know. Did you get a chance to have a look over the questions, but.

Speaker 1

I had a quick look I didn't like write anything down, but I have an idea of them, yeah.

Speaker 2

Yeah, that's perfect. So I suppose I just wanted to start off by asking you your background and just kind of briefly your kind of knowledge on Transformers and music generation and just have an idea.

Speaker 1

Yeah, of course. So I did. So I'm X. You know, I did a masters at a French school. My masters wasn't like general engineering is what you call an engineering degree in. France. So no specialisation per say. It's pretty generic, but I also have. A masters in acoustics from the University of Adelaide and a master in like minor and data. Science slash analytics from. At that same university, which is in Australia. In terms of Transformers, my background and like Transformers and music generation, I actually didn't have any before the internship that I'm currently doing. Apart from like personal interests, just like reading up on papers because what I'm gearing to do in September is starting a PhD. On pretty much that subject, so it's. Just it was personal curiosity up till then and right now I'm working on it and right now what I'm doing is I've been at X for the past six months and I've been working on symbolic music generation. I imagine I'll like tackle the difference later on in the questions, but. Conditioned on audio, so that's pretty much brought me up to speed on like the. State of the. Art with Transformers specifically used for music generation. So yeah, I don't know if. That answers your question.

Speaker 2

That does answer my question. That's brilliant. Yeah. So that's great. So you you have the experience of working on that with Sony, I suppose to kind of start it off, I suppose my first question is kind of what in in your kind of opinion what makes Transformers a suitable approach? For music generation, compared with other methods.

Speaker 1

Yeah. So just kind of briefly put into context, the other methods that exist. Are you specifically talking about symbolic music generation or audio, audio music generation? Symbolic. OK, yeah. So the previous state-of-the-art before Transformers for symbolic was mostly recurrent neural networks. And like LSTM's, TRU's, stuff like that.

Speaker 2

Symbolic symbol.

Speaker 1

Which makes sense because symbolic music is kind of an inherently sequential modality of music. In the sense that it's literally grid aligned sequences of notes, and if you can find a way to represent those notes in a in a tensor, then you can do sequential depending on it. The thing that makes Transformers specifically with self attention. It's very relevant. Is essentially the same reasons that make it relevant with regards to text processing, which is or like large language models, right? Is that the context length is greatly lengthened compared to like RMS and OSTMO'S and self attentions mechanism that not only enhances explainable ality because you can kind of. Figure out which parts of the input influenced the parts of the output. But also controllability, so there's a lot more options for conditioning Transformers or symbolic music generation through Transformers by just appending things to the input sequence, right. And to me those are like the main. Interesting things about Transformers for symbolic.

Speaker 2

Yeah, yeah, yeah, that's brilliant. So kind of jumping over to say like the data set and is there a relationship between training data size and diversity and the quality of generating music by transformer models for polymer? Paraphonic music generation.

Speaker 1

Yes. So as as obviously the there's a lot of parallels with like text right and large language models. It's well known that the more data you have, the more performance they are. The advantage with symbolic music is that the vocabulary size is much, much smaller. I mean it kind of depends on what. Resolution. You limit yourself to if it's 128 notes, it's going to be kind of a nightmare for the model to to learn. But if you limit yourself to 32nd notes, the standard vocabulary size is like 300, something like that compared to multiple 10s of thousands for these models. Which kind of brings down the requirements in terms of data set size, because the less vocabulary you have, the less data samples you need to be exposed to the whole vocabulary. But the so that's for the the size in terms of diversity, the more diverse your data is, the more diverse your outputs are going to be. So for instance, I'm working on the base base generation, so not piano, but specifically base and base is kind of much more repetitive. Piano and it's part of my experiments was noticing that, yeah, the outputs are much more repetitive, but that's just due to the nature of the data that's fed to it. So it just kind of makes sense that if you want a diverse output, you would have a diverse data set, but I don't think that's the quality of the of the output.

Speaker 2

Yeah. Yeah, that's very true. I mean, you can't expect it to have a diverse output if you bring it in. Yeah, that's correct. So I suppose following on for that is how well can transformer models generate diverse and original?

Speaker 1

Yeah, exactly.

Speaker 2

Music and how effective are they at avoiding repetitive patterns and generating novel music ideas?

Speaker 1

Sure. So I think. There's been a pretty interesting research path recently that's been multitrack symbolic music generation, and that's pretty cool because it essentially. The goal is. To generate cool arrangements instead of just piano, which was the first step into symbolic generation and. I am. I think it's good in generating diverse. Outputs because the data sets that we have in terms of, like MIDI, which is the modality that's used to represent symbolic music, are pretty huge now. So it's getting like current models are getting relatively good. At outputting diverse generations specifically because there are more and more sampling techniques that are coming out to enhance that diversity by just like. In putting more quote UN quote selective randomness into the generations, so like smart, smart randomness. Essentially, I don't think it's very good at generating original like new ideas, essentially because. This is kind of the bottleneck with language models and also symbolic music generation models. That they can only pretty much reproduce what they've already seen, right? And without a like reinforcement learning layer on top of that, they can't really learn new things and learn to produce new things. So in terms of novel, I don't think it's very good at that, but the part. One of the question that kind of intrigued me when I. First read it. Is that repetitive patterns in music are mostly a good thing if you're if you're. Generation model learns to produce repetitive patterns over the course of a song, so it could be like 2 verses or two choruses and learn to, you know, it's a it that's actually a desired trait and it's something that's been specifically studied and like tried to be. So just, yeah, it's it's avoiding repetitive patterns isn't really a a subject here. It's more trying to get the models to be musical and be repetitive when it needs to.

Speaker 2

And is there a balance then between 4 repetitive patterns like like you're saying? Like it's fine to come up as a course or? And something in that kind of song context but. Say, if you're you're getting the same kind of the last couple of notes over and over again in that kind of context. So is there, would you say there's a balance towards repetitive patterns and is there a way of combating like the the kind of, if it was mundane and kind of boring repetitiveness? Do you understand me?

Speaker 1

Yeah. Gotcha. Yeah, I think that boils down essentially to. Showing more data to the model and the sampling techniques that I mentioned previously. Which there's an interesting paper that's called. Theme music transformer which kind of tackles that explicitly, which is an interesting read and I think it has an answer to some of the like subtilities or sorry subtleties of that question.

Speaker 2

Yeah, yeah, that's brilliant. Actually. I'll have a look. At that. So then going on to my following question in the context of music generation, what are the key challenges associated with capturing long term dependencies using the transformer you were speaking about that it has improved? From R&N's, but is there still a challenge there?

Speaker 1

I would say the challenge is essentially the same as in. Most like language tasks. The more context you have is always better and you have to kind of avoid. That context is vanishing overtime, like the model still has to try to pay attention to the tokens that are as far as possible from where you are in the zone. Some of the dances have been, so the goal is essentially to have the longest contacts possible, right? And that's been done by the linear transformer which. Kind of. Squares the memory requirements which allow us to basically square the input context, but. Now, I don't think it's as much. Of an issue. The issue is more. Having expressivity rather than seeing as having the biggest receptive field possible for the model so. I would say the key challenges are they're not solved, obviously, but they're not as much in the forefront in terms of having long range dependencies as much as like having music specific dependencies now.

Speaker 2

Yeah. Yeah, that's a fair point. Sorry, I'll go on and what methods or techniques can be employed to encourage transformer models to produce more diverse and novel music? I think it's something we kind of crossed in the previous question, but is there kind of specific techniques to?

Speaker 1

No worries.

Speaker 2

I know I'm saying avoid repetitiveness, but like I was saying like the more mundane redness is that, do you know of any specific techniques?

Speaker 1

Yeah. So just just to go over again what I what I said in the previous questions, the sampling techniques are pretty important because I mean you, you you probably know how Transformers work and I. Assume you do. But essentially the if you just select the the maximum probability. Broken every single time. It's going to be pretty boring because it's just gonna, you know. So, like, I don't know, with C or the most common note that occurs every single time after a given sequence of notes. Which is going to produce boring and mundane patterns if you implement smart like some random sampling which could be nucleus sampling. There's temperature sampling. These are all techniques that come from language. So that's one way to avoid repetitiveness and kind of encourage randomness and diversity in the outputs. One thing that I read very recently and this is more like a proof of concept than anything else, but it's like very, very recent research is having a Dictionary of. Already produced patterns in the song right and kind of use those produced patterns to influence the rest of the generation. So instead of only paying attention to itself, the the model pays attention. Or instead of only paying attention to the input sequence, the model also pays attention to. Sequences that have been evaluated as having a high motif score or whatever, so that's a very, very recent technique. I don't even know if there's a paper out on it, honestly, but it's.

Speaker 2

Is there a name for that? Sorry.

Speaker

It's an interesting one.

Speaker 1

A name for the. Techniques, yeah. Honestly, I wouldn't know. I would just call it like motif cross attention. If I were to give it a name, but it was literally just around the table talking with other researchers so.

Speaker 2

OK. Yeah, yeah, that's OK. OK. Yeah, that's good. That's actually that's very interesting because I suppose that's the point that there, there is motifs in the piece or you want repetitiveness. But yeah, like that getting that that balance. So, umm, how do you pre process or present the music data to be compatible with the model? So is there specific things that to look out for when you're getting your raw data that you need to get it so that it's actually you're able to work with it?

Speaker 1

Sure. So with kind of traditional music, symbolic music, data sets. I wouldn't. There's not much cleaning to be done per se, because they're pretty much already pre cleaned. I would say that that also depends on your use case, like for instance, if you're specific. If you don't want your model. To be learning like. 64th note resolutions and we want it to be rather simplistic in terms of what it's trying to do. You would obviously clean all the notes that are below 64th note timing. The main thing I mean the main thing to make the data compatible with the model is the data is usually in MIDI format, which is a. A. A protocol that allows essentially computers to read music right, and then you would have to tokenize that MIDI data. So MIDI data is a sequence of events which contain the information of. Start like start time and duration and you want to turn that information into essentially tokens within a vocabulary. So you have to expose a tokenizer to the entirety of the data set so it learns all the possible tokens that there are. And then you decompose every single MIDI nodes. Into is 3 attributes which you put like in a sequence. So a note would be instead of being. Kind of vertical tuple of attributes. It would be the sequence, pitch, duration or pitch start duration right. And then essentially you just use an embedding layer at the start of your model to turn every single token into embeddings that can go into the transform. But that's the same As for language models, so it's not. Not anything new. The main thing that you. Have to ensure is that your tokenizer. Learns you know, to transform this the the sequences of vertical tuples into a sequence of attributes, like a linear sequence, essentially.

Speaker 2

OK. And and when you were talking about the 64 note timing, can you kind of maybe talk about that bit more? What do you mean by that is this? That they're just. Longer timings than just like I'm. I'm thinking of four, four and these kind of timings. So when you're saying?

Speaker 1

Oh yeah, sure. Sure. So sorry, can can I just ask you like background in music theory and things like that, no.

Speaker 2

So I didn't. I did piano when I was growing up, so I've gone to in Ireland. It's great up to Grade 8, but yeah, that was a few years ago. So I know roughly the theory, but yeah, I'm. I'm not a musician. We'll just say that.

Speaker 1

OK, so if if I say like quarter notes, 8th notes you you have, you know what I'm talking about, OK.

Speaker 2

Yeah, I know a thing. Yeah.

Speaker 1

The the thing with the 64th notes and this is kind of a specific example to what I'm doing. So as I said, I'm doing base. And baselines are usually much more simple than what happens in Kano, right? And 64th notes which are. These really, really, really. Short notes usually don't occur, and if they do in my data set, it's usually a mistake, right? And that's why I said depending on your use case. And your knowledge of what you're trying to generate, which in my case I know that base usually doesn't have anything below 16th note or it's really really fast. I would just, you know, philtre through the notes and say, well, this is a 64th note. It doesn't need to be here because it's not pertinent. To my. Application and that's really just what I meant because the the smaller the notes are, the more. Kind of intricate the sequences and the more difficult it is for the model to learn it specifically, if those notes are very rare, which 16th notes wouldn't be that rare for piano, for instance, but or for for violin. Or anything that plays fast essentially, but for a base, which is usually typically slower and much more locked in with the timing, it's not really useful to have them in. The data set.

Speaker 2

Alright, OK. Yeah, because I was thinking that's very because I'm thinking of 6 fours and 64. So I was thinking that was that's very fast.

Speaker 1

Of the. Like the the time signature you mean?

Speaker 2

Yeah. Is that what you mean by 64 note?

Speaker 1

Not a 64th note is like eight quarter of 1/4 note. So it's like you take 1/4 note, divide it by 4. It's a 16th. Note like that by to 64th note.

Speaker 2

OK. OK. That's great.

Speaker 1

It's just like super, super, super. Fast notes, essentially.

Speaker 2

Yeah, yeah. Yes, probably more in jazz and those kind. Of songs. Yeah, that's really interesting. So. For hyperparameters, is there kind of like a particular one that is mostly used? Or is there any specific best practise with how you'd go about using? Hyperparameters to improve your model.

Speaker 1

Sure. So there are some pretty well existing and explore guidelines for training Transformers. Essentially the bigger your batch size is, the better it is. The also, the longer the sequence length is, the better it is like there are some things where essentially you just Max it out until until your GPU memory can't take it anymore. So that's pretty like. No, no real questions there. I would say the biggest. The biggest hyperparameter that has an influence would be the learning rate, because Transformers can be very fickle with learning rate, and specifically if you don't like warm them up with like a linear linear learning rate in the first phase and then decreasing can often diverge. So that's the thing that I paid the most. Attention to. And it's essentially by trial and error, right? Because these models are so big and takes so long to train that you can't really do a grid search. On the hyperparameters, for instance, the model that I'm training right now has been training for four days, so I can't really expect to, you know, just do 16 searches on the best hyperparameter and it takes four months. So honestly, it's kind of. Following the guidelines that have already been established and then for the learning rate, which is kind of case dependence, you just. Try a small learning rate and if you see that it's or rather, you would. Yeah, that's that's the way you do. You would do you start off with like the standard learning rate which is 10 to the -4 and then if you see that it's diverging, you just go down to 10 to the -5 and you just kind of, you know search by yourself. But it's not something you can automate very easily.

Speaker 2

So are you saying that you would avoid using hyperparameters just because it's it's so computationally expensive?

Speaker 1

What do you need?

Speaker 2

Sorry, sorry. Like grid searcher. What am I saying? Yeah, like using grid search to improve your model.

Speaker 1

Yeah, I would for sure avoid that.

Speaker 2

Hi playing through tuning, that's what I'm trying to say so.

Speaker 1

Right. Exactly. Yeah. No, there's really no points or like. Gain in doing hyperparameter tuning because you're going to spend like six months tuning one model, and if it doesn't. Work out. Then you have to spend six months. Tuning another so. It's essentially user defaults used the guidelines that have been established for Transformers and then for learning rate. You can fiddle around with it until it converges pretty well essentially. But that's very specific to these large models, right? If you're doing like a. Binary. Sorry, decision tree or whatever.

Speaker

You can.

Speaker 1

For sure do some grid searches but. For Transformers it's. Useless, except if you're like Google or Facebook, essentially.

Speaker 2

Yeah, yeah, yeah. That's what I was. So that's part of the reason why I was asking is cause. Yeah, unless you have some massive computers. Uh, it's tricky. So that kind of like leads into kind of, how do you handle the large scale training large scale? Is there like kind of more ways of making it more efficient or like you were saying it took you four days to to train the model you're working on at the moment. Is there certain things that you're looking out for to? Kind of combat that.

Speaker 1

Yeah. So actually I managed to take down training from 2 weeks to four days, mostly by paying very much attention to the data. Because what I'm working with is like 8 GPUs and the model isn't that big like it's 100 million parameters, but it's compared to like current modern language models. It's honestly not that big so. The the computing time in terms of like forward and back propagation wasn't very much of A worry, so you have to you have to make your loading your data loading as efficient as possible and you have to have those GPUs running full time and not waiting for the data to arrive. And so I wouldn't say my model is large scale yet. I would say it's mostly like medium scale compared to what's being done today, but. I would assume for large scale models it's just once you've ensured that the data loading is as efficient as possible. It's just adding more GPUs. You can't really do anything else apart from that.

Speaker 2

OK, that's perfect. And so I suppose this goes on to kind. Of how do? You evaluate how effective your model has been. UM. Is there kind of metrics that you use to evaluate whether you know it's working better than? We'll say an R&N or or another model that you've been using, sure.

Speaker 1

The in terms of objective metrics, there aren't actually a lot like they're the the loss function obviously, which is cross entropy has been kind of the reference for symbolic music generation so far in terms of objective metrics. I don't know them off the top of my head, but I know that recent papers have like tried to implement timing coherence metrics and like harmonic coherence metrics, which essentially evaluate if the model is correct with regards to its own generation in terms of like timing. And Harmony specific to my use case, the. Because I'm generating based on audio, right? The goal is also to evaluate if the model is coherent timing wise with regards to the audio and if it's coherent harmonic wise with regards to the audio. So it's kind of like a cross correlation matrix between the scale of the audio and the scale of the. Of the generation or like the chromatograms, I don't know if you know what that is, but. It's like a. Scale representation of the pitches that are present in the audio and the output.

Speaker 2

OK.

Speaker 1

So essentially you have, you know, the 88 notes which are brought back to the 12 notes that exist. So like just the chromatic scale and then you sum the energy of the audio with regards to those 12 notes that gives you, like a vector of the energy. Present for each pitch in the audio you do the same for the generation. And you just kind of cross correlate those and the higher it is the better it is, but mostly the to evaluate the effectiveness, it's honestly mostly subjective studies. So you just ask people, it's it's honestly it's obviously a bit more complicated than just asking people if it's better. But you give them different generations and you ask them to evaluate them on a scale from 1:00 to 5:00, and then you do a statistics study to determine whether or not within a given P value range, the model that you've proposed is better at like tricking humans essentially into believing that it's a real.

Speaker 2

Yeah, that's that's actually a good point. So like, would you survey people and like like how would you set up that kind of survey to make sure it's statistically significant?

Speaker 1

So I would generate a bunch of. I would use my model to generate a tonne of stuff, use previous models, generate tonne of stuff and then. Just I would make a web app or something or whatever, not necessarily really important. What format it is, but show the three generations to people and then ask them to, you know, choose between the three. Or you could also show one generation at a time and ask them to rate on a scale from 1:00 to 5:00 or 1:00 to 10:00. How realistic the generation is. You could also ask the question it has. It's been played by human or is it AI generated? You know? So it's kind of like. One all versus all one versus all or one versus one right. And from those you can kind of evaluate the proficiency of your model, but it really depends on what you're trying to evaluate like. The fact that. Has it been played by human or not is? First of all, very hard to achieve and the second of all, it's not necessarily of interest. The interest is more comparing to previous models than seeing that you've improved upon those.

Speaker 2

So was it just basically asking them, like, do you prefer sample A or sample B? And like is it because it's it has good repetitive and it you know those kind of questions?

Speaker 1

Yeah, yeah, it could be those kind of questions. I mean, I've never. Set up a study. Like that myself, but I for having participated in a few of them. I imagine it's that those kinds of questions, yeah.

Speaker 2

Yeah, that's perfect. So would be more commonplace, but definitely in your company when you're looking at those kind of models to do those kind of surveys. To see if that's. If the the music is working for you. Are the models are.

Speaker 1

Yeah, yeah. And then of course, there's, like, a first kind of screening of. The researcher himself, who just kind of listens to a few generated samples and is like, OK, well, this really. Is not good. At all, so have to train another model, but once you get past that first screening phase, then you go into the objective metrics which are. Used to compare. The objectivity of if the model is better to the previous ones. And then you go to the subjective. To figure artistically, it makes sense, right?

Speaker 2

Yeah. OK. That's great. Yes, so I suppose my next question is just like another kind of you were saying that the the only metrics you you kind of use is like the loss function and then comparing the the cross correlation matrix. So I suppose the other question I had was how would you? Is there any specific metrics for diversity or originality? I know you're saying there is not really much originality with the transformer, but I suppose diversity and. Yeah, I just wasn't asking the same question again, but.

Honestly, I I wouldn't. I don't know of any objective metrics. I'm sure there's some papers that address it, but I haven't come across them yet, so I wouldn't know. To answer that.

Speaker 2

Yeah, that's perfect. Is there anything like? So I suppose we were talking about before that with jazz or something like that, you would have a lot faster notes. And so when you have a type of music style and we'll say we're using the piano and there's lots of different genres, you can go down with that. Is there any way that you can using the model? Kind of impact the characteristic of the music that comes out or is it dependent on what you're putting in or yeah.

Speaker 1

Sure. So it is dependent on what you're putting in to the primer that you're using, but you can also generate from scratch, which is just like giving it a START token and then it generates from there. There have. There have been instances of influencing the generation, so the the paper that I talked about earlier, the theme music transformer influences it with a motif like a recurring motif, write the song. The paper multitrack music transformer, which is by. So there are two no, sorry, it's the multitrack music machine or I don't remember, but it's by Jeff ENCE, which is JE double SENS. And that actually uses a node density, a slider to influence the generation. And that's it's essentially conditioning. So at during the training they use an embedding layer to embed the node density of the sequence, which they compute, causing state preprocessing state. Sorry, which allows them to later on kind of specify. Zero to 1 hell dense. They want the sequence to be. UM and this is. To me, that's the most that's been done, because before it was just like a temperature slider and temperature is. Essentially, the randomness of the sequence which isn't very musically relevant like no one would want to control that from a production standpoint, right? But yeah, you can you can influence using note density. You can influence using key which is a, but that's just like appending a a key token at the beginning of the sequence. UM, you can condition using chords, but that's the same thing. You just can append the chord sequence to the beginning. That you can check out. The paper pop music transformer for. That I think they use chords. Yeah. I mean, there are ways to influence it, but The thing is. Musically relevant objective metrics are so hard to compute. Because they're often subjective, like it's hard to have a happiness slider or a or a. The harshness slider because those are very subjective, subjective values, so it's it's kind of dependent on one you can what you can actually compute during the preprocessing to influence it, right.

Speaker 2

Yeah, definitely. I am. Is there any biases or limitations to in training the data to ensure the generation of music remains unbiased or respectful of the various cultures? So I suppose. It's a bit of an ethical question of, like the ethics behind the music. And yeah, is there any way of kind of combating that?

Speaker 1

Honestly, the data sets that we have right now are very Western centric, but that's kind of a limitation to all musical data because musical traditions are. Rarely or not rarely, but like scarcely documented most of the time, and obtaining audio of those musical traditions is hard. So obtaining MIDI is even harder. So if there is such MIDI, it's probably like .05% of the data set. So as of right now, I would say that's a big flaw in the field of music generation as a whole, but specifically symbolic music generation is that we're very, very, very constrained to Western classical music or pop, essentially.

Speaker 2

It just.

Speaker 1

There's no, like, real steps to take instead and additionally to like documenting more of these musical traditions.

Speaker 2

I suppose it's just not a bridge that has to be crossed yet because of the limitations of the data sets.

Speaker 1

Yeah, kind of. I would put it. I mean, it's a bridge that has to be crossed, but there's no like tools to cross it yet. I would say essentially.

Speaker 2

OK. Yeah. Yeah. Perfect. So, have you ever encountered unexpected results in challenges during your research that have provided valuable insights into the capabilities and limitations of Transformers for music generation? So any kind of positive accidents, basically positive accidents?

Speaker 1

I've no I have. I've had some negative accidents. The thing that was kind of a A. Yeah, there's been one. So as I said, I train on base right and I was kind of hoping that or I was kind of dreading that the repetitive nature of base was going to be overlooked and there wouldn't be within the same song, like enough repetitions, because honestly, based sometimes there's like. Doom. Doom. Yeah, which is fine with what base does, but because there's been. Because there is no specific. Mechanism to ensure that repetition. I was afraid that I would have to implement one and when trading my first iterations of the model which were not conditioned by audio yet, I found that actually the the repetitive the repetitiveness of the bass was very well modelled in and of itself and even with. Like a bar would like have 8 notes of the root of that bar and then it would go up to the next route, you know? And I was expecting it to have seven or six or not necessarily follow that pattern very well and kind of not model the. The the nature of the base, which is kind of to under underline the chords of a. Bar and it did. So that was, I wouldn't say it was a happy accident, but it was more like a oh, cool that. Works on its own, which is, which is funny and interesting.

Speaker 2

Yeah, yeah, it's nice when it works out that way. It's not always that way. Yeah. So just, uh, my second kind of. Or ethical question, considering the potential use of AI generation music in various contexts, as what ethical considerations should be taken into account to ensure responsible and respectful usage of the technology.

Speaker 1

So this is kind of a a contentious point within research circles. It's in its tangent to the whole debate on can you use copyrighted material? You can use just anything that you find online to. To train. Fortunately, it's not really the. It's not really an issue for symbolic because you usually to have clean MIDI. It has to be the artist that volunteers at, so there's no issue. Usually the artist is consenting to it's data to its data being used. I would say the problem arises when you train on audio specifically, and because there's a real subject and like using audio to condition symbolic music generation, which I'm doing. I'm trying to do right now. It's the same. Issues as using audio in any other training setting, right? So it's a very grey area right now and most labs are tending towards the dark grey of that grey area and that they're using pretty much anything they can get their. Hands on but the. The true ethical path would obviously be to have artists agree and consent to their data being used for training and to have a full disclaimer on what's being going to be used for. Maybe even watermarking? I don't know. I don't know if you are. Aware of like neural watermarking, but it's kind of a mechanism that you put into the models so that you know that that generation comes from or has taken inspiration from a certain piece of music, right and.

Speaker 2

OK, I've never heard. Of that before I've. Seen watermarking on a piece of paper, but maybe not like an audio file.

Speaker 1

Yeah. So it's it's really interesting. It's just like training the model to input like imperceptible noise or like artefacts into the audio that once they pass through, once they pass that artifacted watermark on your 3D model, again, it outputs either an ID with the model name or the model version that's being used to generate, like the name of the artist or. Things like that, right. So it's that's a step, but it's also a pretty. Novel research domain, so it's hard to combine that with all your generation, which is also a novel research domain and have those be ethically responsible together at the same time. So. It's it's a conundrum, but the best thing would be to have. Yeah, those. Consents and, like ethical watermarking practises, be put into place.

Speaker 2

Yeah, it's interesting that that is actually possible. It's there, but it's just like you're saying it's very novel. It's very niche and very new. But it's great that it, you know, there is a beginning somewhere anyway. And is there any like particular resources or places that you kind of reference when you're researching yourself that you think would be helpful towards the research I'm doing at the moment?

Speaker 1

Yeah, I can. I can send you my Mendeley library, which for for this internship, which has basically I hope every single paper on symbolic music generation in the past. Like 5 years. Hopefully so I can I can send you that for sure. And other than that, I think that. Because of the like blatant similarities with language models and language processing, it's also very interesting to kind of read on those and see if there are any any transferable ideas of that and then used yet because I'm I'm sure. There are like. There are there have to be. Things that people haven't thought of using. The symbolic user. Generation. But it's just because it's hidden in like the massive mountain. Those language model research is being done right now, and it's just waiting to be found.

Speaker 2

Yeah, definitely. Yeah. I think this space is is only going to grow even more. If you are going to do this research yourself, is there anything else that I haven't asked that might be of interest for me? To know or. Yeah. How would you go about it?

Speaker 1

So the the research that you're doing is, are you actually training any models or are you just seeking to learn more about the like symbolic usage generation?

Speaker 2

Well, I'm. I'm. Going to be making a a model and I'm going to be using different data sets to see like how what kind of output I'm getting. Is there a difference if I'm using smaller data sets or more diverse ones like something very basic that? And then obviously trying to improve the repetitiveness, hence my multiple questions on that and yeah, and just comparing how the the transformer works with that and seeing if if it does work, is it better than the usual RN.

Speaker 1

OK. Is there anything else that would add or consider not necessarily like? Specifically related to the research flow, strongly advise you to not implement the models from scratch yourself, which is basically like ripped them off, or GitHub rebuilt. That should be open source, of course, but it's I complicated my life at the beginning of the internship with that, and I don't think that is something that was hard to do.

Speaker 2

Have GitHub repos that you would suggest to have a look.

Speaker 1

Yes, sure. Do you use Pytorch or Tensorflow?

Speaker 2

I think I'm going to veer towards pie torch just cause I've looked into it and I think there's been positive kind of people are looking at Pytorch more I think, and that seems to be the.

Speaker 1

Yeah, I prefer torch. Yeah, I mean. You can. I'll. I'll send you a couple and then every single paper usually has their or at least every research paper has their code relatively open source, so you can just control F GitHub in the paper and usually a link pops. Up, yeah. And to consider, I would say. Pay attention to. Evaluate the quality of the data yourself before training on it, right? So kind of listening through like know your data essentially is what I'm saying. Listen through it, do a lot of analytics on it before just diving into the training to figure out if there's like any cleaning to do. Do anything that you don't need for. Your specific use.

Speaker 2

Is there specific things that I should be listening out for? Like if I am listening to a piece of music, is there certain key things that I'm like, right? OK. I need to clean that and. If so, working techniques would you suggest like you know, roughly.

Speaker 1

Are you are you going to be? Doing essentially keno or.

Speaker 2

What? Sorry, Kano. I.

Speaker 1

Are you going to? Piano or you going to be training only. On piano.

Speaker 2

Piano, yeah.

Speaker 1

OK. Not nothing to listen for specifically. I think that an interesting, honestly, an interesting thing that you could do and that hasn't been done yet, I think is modelling triplets. So if you if you want to look into. That kind of like. Because I think it would be an interesting contribution and it's not. I don't think that hard to set up. That's just pre processing to kind of identify them in the music, which hasn't been done yet. Yeah, that's that's to be an interesting path to follow up.

Speaker 2

UM. That's interesting. I haven't thought about. Yeah. And is there like anywhere like do you search for open source data sets and is there anywhere in particular that you kind of find your data, your data sets or you know, I know when you're going through papers they have like their data sets and they have them, you can look them that?

Speaker

OK.

Speaker 2

Is that the way you go about this? Or how else?

Speaker 1

I use. I use papers. Yes, I also use Zenodo, which is general. That's CENODO, which has like a tonne of data sets. You can probably search up MIDI and find something. I also use papers with code. I think most. Importantly, the land. I'll just pull this up for you real quick. The Izmir, which is International Society for music information retrieval, which kind of has regrouped all that, all those papers on symbolic news and generation in the past years, has a resources page with all the data sets that exist. And probably if you look up MIDI or like symbolic in that data set in the in that page that you'll come up with everything that exists already, so.

Speaker 2

What's the name of that page again? Sorry.

Speaker 1

I'm just sending the link in the chat right.

Speaker 2

Now. Oh, grace. That's really hard.

Speaker 1

So yeah, if I just control F MIDI, there's like 2929 occurrences, so you can probably find everything you need.

### Participant 2

Francesco Foscarin.m4a

Transcript

Speaker 2

And I can refer back to it later on and maybe reference it in my research.

Speaker 1

Yeah, it's OK. It's OK. So this is so they interviews themselves as part of your. I mean, part of your pieces. So you want to ask opinions about this topics?

Speaker 2

Yeah, exactly. And just to kind of get more more of an idea of what I'm trying to trying to do, if there's anything I've missed or am I suppose it's kind of there's only so much you can take from reading and if somebody's actually done it first hand, it's, you know.

Speaker 1

OK. Yeah, sure. Did someone in your university working with this?

Speaker 2

Is invaluable. No, it's just something I've decided to do myself.

Speaker 1

OK. Yeah. So it's definitely good to speak with someone, yeah.

Speaker 2

Yeah, but no. Yeah, my supervisor only knows stuff about finance, so he doesn't got a clue what I'm doing.

Speaker 1

OK, OK. Yes, yes.

Speaker 2

I am. Yeah. So that's great. I don't know if you've got a chance to have a look. At the questions.

Speaker 1

Yeah, just very briefly. But yeah, yeah. So I'm not. I'm not prepared also this very comprehensive list of questions that you prepared. So this is nice.

Speaker 2

That's perfect. Yeah. No, that's. That's great. And we'll we'll go through them anyway. And you know there's no worries if.

Speaker 1

Yeah, what do? Probably not, yeah.

Speaker 2

We can just go on to the next or whatever and but yeah, so I'll, I'll start from the top and just kind of get an idea of do you want to just introduce yourself and maybe get a bit of your background and?

Speaker 1

Yeah, sure, sure.

Speaker 2

Kind of, yeah.

Speaker 1

Sure. Yes. So I'm now at this moment I'm in the office in Vienna. I'm working for the university in Austria. As a postdoc, I completed my PhD in Paris, so I I work also for my master working with field of. Music information retrieval or music information research. Is called. Which kind of group all kind of research. So it's not it's not only generation and in my case I actually never did paper for example specifically about generation, but there are many other stuff like analysis. You know, it can be people's behaviours and other search. I mean basically if you're not generating or analysing something in general or starting something. But usually yes with with. With operates with computer systems with machine learning. Deep learning? Yeah. So this was always my field. I've switched a bit more to deep learning field in the last two years when I was here before it was other stuff also like. Or algorithmic stuff and more. So database and application. But yeah, I I prefer this and of course also this deep learning is very research right now, so it's nice that you can get different. I mean many, many new research are coming out. So there are many new inputs. This kind of research.

Speaker 2

Yeah, prediction is, yeah, particularly in the arts as well, like in art and in music, it seems to be kind of everywhere.

Speaker 1

Which is nice. Yeah, exactly. Yeah, yeah.

Speaker 2

Yeah, that's perfect.

Speaker 1

Yeah, I mean, you could go inside more and more ethical question and stuff. Of course, people have different opinions, but I don't think it's your. It is your concern for now.

Speaker 2

Yeah. Well, we we might touch on that a little bit later on, but yeah, so I suppose I suppose my first kind of question after that is like what like do you have you worked with transformer models and?

Speaker 1

Yeah, exactly.

Speaker 2

And what makes it suitable for music generation as opposed to other models that you might? Have worked with.

Speaker 1

Yeah. So I will pretend. And yeah, I mean nothing. Nothing is specific in this. Basically. I mean from from these fields of art and stuff, usually people. I don't have the. Time or, you know, I mean there are just much less research in this music field or in other with respect, for example, to natural language processing or image processing. So very, very often, I mean almost exclusively. People just steal these kind of models from from this. For example, in this case this was a natural language processing model. And then people decide, OK, let's start to use it. So basically what it makes good for music, it is the same thing that make it good for language or. Image so it is that that you can so so before this what what was the alternative maybe some kind of recurrent neural network model? I suppose this was the other alternative. Or yeah mostly this. So the problem back then was that you had to encode all the. If you have an input sequence that you want to refer to, you have to basically resume everything in in one single. State and this can be complicated if you have a long sequence, because something in the past will be forgotten and information will be a bit crashed. The battery to this one will transform any attention based transformer or self attention based. They can look back or forward at these different tokens and. Extract the relevant information. So basically they they are they. They should be better at at modelling last long distance relationship. Basically that's what it should be. Sorry one second.

Speaker 2

Yeah, that's OK. And yeah, that's perfect. So when I was reading up on it, there was a lot of different transformer models and they're kind of targeting different aspects of music. And I suppose in my research, I'm trying to. You know, obviously having reputation in music is good, but trying to minimise the mundane repetition, like if you're just getting the same kind of view notes every so often, it's not very interesting to listen.

Speaker 1

Yeah, yeah, that's correct.

Speaker 2

To so that and trying to get more diverse music and good long term dependency. Do you? From your research of different transformer models, is there one that's best fit for that? Or is there you know the main transformer? I know it's very much what you're trying to do. It depends on which one you would use, but in your opinion, is there kind of 1 you would go for?

Speaker 1

Yeah. So I don't know what you mean, but different performance models, you mean this this popular architecture with fixed size or like different attentions, I don't know.

Speaker 2

So you know where there's like, there's the vanilla transformer, the XL transformer motif scene, those kind of.

Speaker 1

OK. Yeah, yeah. OK, let's see. See. So I mean it's it's always a bit tricky in music basically because we don't have as many data as the other fields have. So usually what is done is you keep something smaller. I mean you you. I mean it depends what? So if you're going from some something where you have a lot of data, some people are now doing some research with also huge amount of data. But yeah, I mean I there is no. I mean, this is something that we need to be done with some studies and I'm I'm not aware of a study. That is trying. To find what will be the best option so people are a bit going randomly, I will say this is also something that is going on in the planning all the time. So no, I mean I I will not have a suggestion. This will depend on. On how many much? Many. Data you have. So you can. Also have a generative model with something very small. It will work then. In this case I would suggest not not using any of this architecture, just just having you know a couple of layers, even 2 layers. I work in the past. Stuck on top of each. Other and of course, if you start to have some huge data sets, you can. Go for this bigger one. There is no. Yeah, there is not a clear choice. Also, maybe there is no. I mean, these models are actually particularly good because they can somehow adapt themselves. So I'm I'm also you you probably. If people will try to, you know like make a study to show what will be the best. I'm not sure they will get a specific answer probably something just random because you know the model will basically adapt in some way This is why there are. So many. Also, components of this transformer where we don't know exactly what we should use, you know, just yeah, people have done this. Then they publish a paper that this is bit better and then other people actually show that for them something else is a bit better and you never have a clear answer. Just just because they can just. Somehow up to well, so even if something is not very good insight even. If you have some. Bad data or or some part is not very optimising. We still work out finally. So for this I don't know. Yeah.

Speaker 2

Yeah, that's that's a fair enough answer. OK. So moving on to kind of like the data set like you were talking there is, is there a relationship between the training data size and diversity and the quality of the generated music by the transformer?

Speaker 1

Yeah, I will need to introduce one missing question here because I think it's very important. So what what kind of data you want this is, this is. Important to set, I will say there are two. Main branch of research with generation models.

Speaker

Now. Mm-hmm.

Speaker 1

So one is audio data, I mean. This is music. You just have an audio file and you want to. Generate an audio file. And the other is what we call symbolic data. I don't know if you're familiar. With this or.

Speaker 2

Yeah. So in my research, I'll be going down the symbolic music route, yeah.

Speaker 1

OK, OK. And then the other difference will be what kind of symbolic, so symbolic data there are maybe. I guess you're familiar, maybe with media files or this kind of. Stuff so MIDI files are ready. So let's say that these MIDI files there can be a lot of different stuff. And there are other more structure files like musical score for example. If you want like scores like you know musical score like sheet music with notes and something that is readable by people. And so this is 1 and the other is like, OK, what if even if you have a MIDI file with a MIDI file, you can encode, for example, a score, some simplified version of this musical score, or you can encode the performance so you can have people playing on the piano and being recorded, so they will be two very different kinds of data. Because if you have a musical score, you can expect like almost, that will happen at the same time to be. Aligned together, like starting exactly at the same time. Well, if you record the player playing, even if you still need file, you know, maybe you play something. I suppose if you try to to play 3 notes at the same time, you will not do this. You know there will be some small time deviation inside, there will not be any concept of. Of power lines or time signature or this kind of. Stuff that you will have instead if you target. So what we call them is like I call them quantized MIDI file, where you where you know all the nodes are at the same time in some specific bins, and you don't have so many temporal position that you can have because maybe you have 8 nodes and maybe. You have 16. Nodes or triplets, but that's all basically and you have some specific time points where the nodes can happen. That depends on this, or if you're targeting performances instead. Basically, this is a continuous space because the player can slow down, can speed up, it can have a node. That is, some milliseconds after the other. You know, so these are very two very different kind of data and. People that do so, I guess if you're reading this depends with median music. Most of them they expect this kind of contest data because these media files were maybe produced by hand by someone on a on a this toe. So they were typed. They are precisely quantized, but there are also. I don't know the some data sets like. I don't know the master data set for example. Which is just people playing piano being recorded, then this is a very different kind of data because you have instead all this time deviation. And you need to handle timing in a different way.

Speaker 2

Yeah. So.

Speaker 1

I I don't know how familiar you are with this this difference.

Speaker 2

So are you saying there's? So what the is? Are you saying that with MIDI files? Generally they're very accurate on when the note is going to be played, whereas if it's a person that's playing it with an audio file, there might be slight differences because they're human and they don't. They have their own flair to how they play the song.

Speaker 1

I mean, this is true. Yes. And and even more so. I I will you. I will really see them. I mean it's it's not me this it is pretty clear like you really need to see them as to very different kind of of of. Of files, so it's like. Let me think. It's like if you have a person speaking. OK, you have all the Inflexion. The person is speeding up, slowing down. Even though this is not so relevant in speaking. But or you have like a robotic voice speaking something. And in music this is even more extinct because if you so you can imagine you have like a I don't know, drum machine something that's that is very precise or or something this so you will you will know and if you have 8 nodes that is. Played by this time machine or by this software, every every nodes will have exactly the same. Duration. So for example if you want to input this kind of data in a transformer to generative you will maybe have some token that say OK, this is an 8. Note this is a 16 node. And so let's let's divide the measure. I don't know. In in 12 parts. And so let's divide each quarter. Notice 12 parts and and save something like this. So we wait this kind of parts and something like this. So this is what you can. But this these files are like very robotic they they will not mimic someone that is playing, they will not have any expressivity. So this is something that you that you can do and if you see this this very big data sets, that's what they do because they take these MIDI files from someone that just input them in. And basically they will probably just have quarter note divided maybe 16 notes, maybe some triplet and then that's all basically while if you take data set of people playing, you will not know where the measures start and where the measure end because people playing don't. Specify where the measure is. And you will not have a. Note that is you will not have an eighth note that is the half. Of a. Quarter note you will. Have something more, something less each quarter. Note in the piece will be played in. A different speed. Different duration people will play notes that are quarter notes shorter because they need to lift the finger. For playing another notch or longer. If you're playing legato, so notes with overlap. And you have all this kind of mess inside your data, let's say.

Speaker 2

So are you saying so? They're. Are you saying that the because? Am I? Am I right in saying that you're saying there's two types of MIDI files? The one that's basically imported like you said, and it's the one that's been made. There's two different types of media, OK.

Speaker 1

Exactly. Yeah, yeah. Exactly. Yeah. You need to be very aware of this when you select your data and. When you select. Your technique because for example, if you read a paper, usually took paper with this what you do with the one that is called quantized. So you have only some specifics. They say OK, actually we look at these MIDI files and we can say that in a measure.

Speaker 2

OK.

Speaker 1

There are maybe, you know if you have a four, four measure you have four quarter notes and if each quarter note you are playing 16 notes, maybe you have like 16 places where the notes can happen in the. Measure. And so they say, OK, so this is a grid that we define and we can have notes happening there. And so we have this. Kind of. For example shift token happening that can shift 1234 of this like smallest divisions. This is what they do. But if you give to these systems unquantized MIDI file that is coming from a performance. So this will definitely not work because first of all you don't know where the measure is, so you don't know where you need to start to. Count these 16 notes and then every 16 notes will have a different duration, so you cannot. No, like you cannot do this shift. Basically you need to have something much finer in order to be able to encode these different. Shifts and tempo deviation that people are playing.

Speaker 2

So are you saying that the shift token is a way of trying to make it slightly less robotic? Is that what you're trying to say?

Speaker 1

No, no. No, that's not it. Wait.

Speaker 2

OK, sorry.

Speaker 1

Let me let me see if I can. Find an example. And just what I'm saying is that this so when you, when you, if you, if you're doing symbolic and you're starting with this, you need to, you need to decide what kind of. OK, can can I share my screen? Maybe yeah.

Speaker 2

Yeah, there's a share screen. If you put in the three dots, there's share screen is an option.

Speaker 1

Maybe you need to enable. Me, because I maybe don't have the rights to do this.

Speaker 2

Uh, can you see it?

Speaker 1

Yeah. No, no, I can, I can. Sorry. OK. I I think it's pretty important. That's why I'm insisting on this, because you, you see. Many times people maybe with not a strong musical experience or whatever, that that get. A bit lost on this and then the research is not really meaningful anymore and it's also not working. Just maybe you know they they they they don't have any result. So this is a musical score now, and this is the performance. So this figure is not ideal for you because what you could have you could have this musical score also 70. Five, you know. But what happened in this MIDI file for? Is that all these nodes that are supposed to be at the same length, they will have the same length. So if in your MIDI file this is, I don't know, 50 milliseconds, whatever. This will also be 50. Milliseconds, yeah. And they will all be the same. And so this is what we call contacts. So there are only some specific times where this. This can happen that is after 50 milliseconds. OK. Instead if you take this one, it's a real performance, it's a MIDI file recorded by some people playing. So you you see some some things that are not. Not supposed to happen in. I mean they are happening because of the performance. So for example these two nodes are coming one after the other in this data. But here they are overlapping just because the player. Will sometimes overlap a bit, so this note is ending after this one started, so they are not sequential anymore. For example, they are not of the same lengths anymore because the player playing something smaller like. Yeah, and they are not. They are not agreed. Let's say there is not greed. And there is also not concept of when this measure will start. So the The thing is, if you're, if you're handling this kind of data in your transformer. You can get away with some very huge simplifications. For example, if you're handling this kind of data. You can say OK. I can assume to have a note every 50 milliseconds, 5050 fifty. Here is 25 years. 25 here is 12 so I can set up some sort of grid which is maximum 25 of of distance between the nodes and you can get away with this and your network will be much smaller. Because you can expect only notes at these positions. And this is this is what the people that that target this quantized or score mini fans are doing instead if you input this kind of data which are also available like the next data set is huge and you can have it and you can train a generative model on this. So you cannot use this simplification. What you need to do here is say OK, I will take a time windows that is very very small. And I will assume that the notes can happen on every. Point of this time grid but this time grid need to be much smaller because they need this time grid need to be needs needs to model these kind of things. And and so you need to have a further, a different, a different model and a different simplification.

Speaker 2

So is that where it's kind of like the event of on off? Is that what?

Speaker 1

Yeah, some people. Yeah. I mean, you can have this this on off in this in in both the presentation let's say. It is just. This is mostly about how you treat the time. So if you if you have this kind of representation, people will have. Let's assume that you have a standard generative transformer. You will have maybe some tokens that that are encoding time, OK. And in this representation you don't need so many. Basically because you will have only some specific points in the measure where this can happen. If you are using this representation, you need much more finer time step. And this will somehow complicate the architecture because you will have much more space on the token that can happen.

Speaker 2

OK.

Speaker 1

OK, you can you can read I mean this this is something some better that we wrote in the past. I can send it to you. Yeah, there's an explanation between these defence. Just just make sure that that. You know. Which one of the two files are you targeting before? Otherwise, otherwise it will become messy, especially if you if you then start to mix The Chew and. Then yeah. I've seen some people doing this and it definitely it cannot work like you need to have. It's like if you are mixing a transformer that generate text and you also have audio files of people speaking in the same architecture. This is definitely not something that will work. You know you need to have. You need to have the same kind of data, let's say.

Speaker 2

Yeah, yeah. That was really helpful.

Speaker 1

Yeah, sorry, this was also long, but I think this was important, yeah.

Speaker 2

Yeah, it is important. Definitely. I think once you have your data right, I think everything else is just easier if you get that wrong, then you're kind of at nothing. And so when you're saying that. So you take it into smaller chunks. If it's a human performance on the MIDI file.

Speaker 1

Yeah. Because you you never know when the. Notes can arrive, yeah.

Speaker 2

And you're saying that? That complicates your architecture. Is it that it's harder for the transformer to predict when the music when the note is going to be played, but is there more to the architecture that you've to add in to try and? Apart from just having shorter.

Speaker 1

Yeah. No, I mean, I I I don't think so there is. It is instead the opposite, because these people it's like about how much you can simplify your problem, let's say. So if you have this course, you can simplify your problem. If the if you have the context MIDI or the scores MIDI, you can simplify your problem much more and and this is what if you read these papers of people. Doing generative they is not very clear many, many times. What simplification they do, they often write. For example they write OK, we assume. To have only. For for music like the time signature is only for four. And this is an idea, this simplification, and then they say OK and we assume to have agreed of 16 positions where the nodes can happen. And they can do this because they have these files where they know where the measure is starting, otherwise they you cannot do. This and then it's just the token space is much smaller than. But in terms of the architecture, you know this this, as I said before this architecture. They are not made for music natively. They're made for something. Else so there is not then I mean I didn't see any. Proposal of these different architecture to render these two kinds of data is just.

Speaker 2

It's just the way you prepare your data really.

Speaker 1

In one case you can simplify much more. That means that you need maybe much less data to have something that makes sense then.

Speaker 2

OK.

Speaker 1

It's like if you're targeting audio, it's even worse. Now you can. It can simplify even less because audio then you will have so many stuff, confounding problems with different instruments and stuff coming inside. Then you will. Need more data? What symbolic you need? Maybe you need maybe less because you already have this concept of nodes. It's kind of a bit. Higher level. And if you have scores instead of performance, this is again a bit more higher level, so then you need. Less data to have something out of the same quality.

Speaker 2

So if you're going with the quantized MIDI file, you would need less.

Speaker 1

I will say yes, you will need less because you can. This is a bit higher level so the network don't need to learn. For example, when a measure is starting because you are providing this information already.

Speaker 2

OK.

Speaker 1

Because the network will know that every. 6 seconds A new measure is starting, or maybe many, many people do this like they have start measure token for example. So you can tell them here. Measure starting if you're using the uncontested MIDI, you don't know when a measure is starting yourself, so of course you cannot keep this to the network. The network will try. We need to figure this out. Somehow, and to do this, we can expect that more data will be needed.

Speaker 2

Yeah. OK. So yeah, there'll be more data and then because you're taking it into smaller chunks as well, it'll be even even bigger.

Speaker 1

Yeah. And the other things may be relevant for you and many. So I said I'm not very expert of this narrative. So I don't know, for example, by heart, this, this, there, there, there were many evaluation. Metrics that were proposed. About for example, even this stuff that you mentioned before, how boring repetition they are and this kind. Of stuff. But as as far as I know, almost all of them, they will only work on contact data. Because they will use this concept of measure and know how each measure is similar to to the next one to come out of metrics. So if you if you are instead producing this and context MIDI you will not have this concept, then you cannot even run this kind of metric for example.

Speaker 2

OK. So I suppose the suggestion is to kind of focus more on the quantized MIDI file.

Speaker 1

I will say this is season as far as I can see it is. I mean, I don't know what kind of what is your generation goal, but yeah.

Speaker 2

Yeah, well, I suppose I I suppose that it it makes sense cause it's more easier to work with, But yeah, I'm trying to yeah not have mundane reputation and more diverse music. And I I see that if it is known by human it, it might be a bit more diverse because there's you're getting notes slightly differently, but. Like you're saying it. It's it's more complicated, it's more confusing for the model.

Speaker 1

Yeah, exactly. So I will, I will probably go, yeah. And and I guess you will also find more paper on this because that's this is more easier than people are going for this option. Yeah. This OK, so so we can go back.

Speaker 2

That's brilliant.

Speaker 1

To the question.

Speaker 2

Yeah, no, that. But I'm glad you you said that, I suppose. Yeah. The next question was more kind of the relationship with the training data size and the diversity.

Speaker 1

Yeah, yeah, yeah. I mean, exactly. I mean, I gain more. Non music specific stuff here. Yeah, I mean everything everything is. A bit mysterious and blurred in this deep learning field, but yeah, I mean, yes, in theory we can expect that more data and more. More data will. Yield better results diversity. I mean this can also could also be a problem like. If you if you have. More diverse data. You will also need. More data because if you have. Different if you if you have, let's say the same size of. The data set, let's say if you have 1,000,000 songs and and only one style, you can probably learn this. Style pretty well. If you have the same size with 10 different styles, you will probably have a mess between these styles, so you will need 10 times more. Data and maybe even. More in order for the model to. Be able to accomplish all these diversity, let's say. So if you if you I mean there are still something that will complicate the work. Of your transformer. Somehow, because you do, you need to be able to handle this this diversity. So just just. Have it if you can have more data. This is good. Yes, you can have the same kind of data. This is not good. This is again what these people do when they say we only have 4 four music. So this is. I don't like this assumption or still because there is so much music in the world. Another time signature. But that's what that people do because they they. Cannot handle them because maybe the data set of 3-4 music or 68 music are not so big with respect to the four four. Then they will not have enough data for the model to be able to handle this kind of difference.

Speaker 2

And four, four is probably. Yeah, I suppose it's simplistic enough for more people to understand.

Speaker 1

Also is nice because you can also always divide in two. You can always have the measure it will. Also have be OK. Right. If you have another sister, you cannot touch the measure because it will be in the middle. Of a bit. Then you need to be added in three parts and this this complicated of course the work of also of the the processing and organisation that people use. This will be more complicated than.

Speaker 2

Yeah. So I I'll jump down to a question and come back up, but how do you pre process and represent the music data to be compatible with the data or with the model? So is there? Specific techniques to preprocess.

Speaker 1

Yeah. So this is very important, especially maybe because media is what, what can it be? So there can be many, many things. So there are there are many different options. So if I will have to to start now with something without really knowing many. I'll probably go with some form of tokenizations, so there is for example there. There are these people from France that made this library that I actually didn't use it. Yet, but it seems to me pretty nice. It's called media talk I. Don't know if you know it. So if you're dealing with quantized music, it's probably fast to use this one, because otherwise this can also be very complicated because symbolic music formats, I mean media is not too bad, but you can have all sorts of problems inside and all sort of exceptions. And if you start to write this yourself. It it can take more time to do this than to to actually write the the transformer architecture and do this experiments you know. So if you can place your work on someone else this will speed your work a lot. So I will say probably, if you're going for standard transformer stuff, probably just go for tokenizations because these papers are there and the people are using it and. Seems nice. And they use some of these famous tokenizes that already exist, that are already implemented, so you can save a lot of time into this. Because it can become very, very tedious with all the possible different kind of MIDI files and different kind of this and this that exist. So. If you if you use something that is already there, is probably faster. I will say it is have been, I mean this is what people do in again this is what people do in language. So that's what people are trying to do just to mimic that is it the best thing for music? I don't think so because these documents are a bit weird like you we need to have a sick. After this and the music is not really sequence because you can have stuff happening at the same time or stuff ending after something else is started. So this is not a sequence anymore, but you can somehow like push this stuff into a sequence model and. And this will work. This is the fastest. I also think there will. There is nice to do some research on. Different ways of doing this, maybe more musical, but this is this is something else here.

Speaker 2

Yeah, they. Yeah, I was talking to somebody. They were saying that they started off with one hot encoding and then they went into a word embedding and. The these sort of things pre process the the data so there's a few different ways to go down.

Speaker 1

Yeah. Yeah, exactly. Because one time you win, we're not encoding. You can. Yeah, I mean these are two different stuff. When you when you use word embeddings, you cannot use them or not encode embeddings. I mean it's it seems to to be nicer to adjust some tokens smaller vocabulary than you use just this impending layers and then you put it them into your network. I will say this is the most standard stuff that. You can do now and be shown in language to work very nice. So yeah, I guess you can go for this. One yeah.

Speaker 2

So yeah, I'll just jump up again and just. So how well can attention based transformer models generate diverse and original polyphonic music? And how effective are they to avoid repetitive patterns and general generation?

Speaker 1

I think, yeah, I think pretty bad at the moment. Unfortunately, there is much research, but there is nothing like even close to this super nice performances that we see on images for example you. Know, yeah. We are very much behind probably. Either for because there is just, I mean the the people that are working on this are very, very tiny fraction. Of the people that are working on images or text, so of course the surface proceeds slower. And the other reason could be that music. It could be much more complex, for example, because you have this. We really have these long term structures and we have. We have structures first that that need to be something, otherwise we will not really like the music that is produced. And this is not only you know about the next measure that is not so repetitive, it's also I mean if you listen to a piece of music, you will. We have. Some some stuff sure that you can expect. Maybe a course that is coming always a bit different and this. Kind of stuff. And I mean this. What I've seen until now is just is we are very far from from having something that just produced this. Some audio owners or some symbolic, so most people maybe are. Mainly focusing on maybe conditioning now on something maybe conditioning on a structure that is given beforehand or conditioning maybe on courts to produce something else. That that. Yeah, I will say, I mean, if if as a call we we are having a music as it will be composed by a human, we are extremely far from this. I mean no one. Will resolve from this. Now from any of these systems that exist.

Speaker 2

Yeah, I I can't remember. It was in the news recently that they finished off a famous classical music song. I don't know. I can't remember who the composer was, but they asked was this. They're like, oh, we found this song and it's been. Finished, but actually it was a I generated.

Speaker 1

They finish and all sound gaming this this one. Yeah. Yeah. So yeah, there is some research on this, but this is a. Bit of a. They always are a bit of. Clickbait headlines. Because what they did in this case or in this whole case that you see on the newspaper, basically they they have an ability that is generating. Stuff and there is a human that is going through. 200 of these generated stuff and decided that this one is the one that you like, and this one is this the one that we put for this couple of measures, and then it generates some other couple of. Measures and then decide so of course this is artificially I mean it's created, but then officially there is, but there is there is the. Human that is selecting. What he likes, so I will not say that this is. 100% I mean it's like, yeah, it's aided let's say it's added composition, but it's not created completely I will say. Created completely something that you can. Just say create and this is and you create something that makes sense. It's very different. From making one measure and then system gives you 20 possibilities and you check you select this one. And then you go. The next and then you say, OK, this is my colours. Then let's generate another verse for this course. This is something a different process. So of course, if you if you are there and you check what is done and you select all the time what you prefer, you can have very good results, but then you're not so far from what will be like. Normal music composing. It's like Jeff suggestions from the. And you pick the one that you like.

Speaker 2

Yeah, it's basically just having another person there, but it just happens to be a computer rather than a human.

Speaker 1

Exactly. Yeah. So this is what? Yeah. So the the very nice songs that you will see generated that they are made in this way.

Speaker 2

OK.

Speaker 1

So there is a musician then that is kind of selecting what he wants from these outputs of the systems, yeah.

Speaker 2

UM. Yeah. So going on to my next question, what methods or techniques can be employed to encourage attention based Transformers models to produce more diverse and novel music with avoiding repetitive patterns? I think we kind of, yeah, touched on it.

Speaker 1

Yeah, yeah. So I'm not, as I said, I'm not. I'm not expert of this, but I can give you some pointers because I had a colleague that is working, so I think so I think what you need to avoid this repetitive patterns and stuff. It's some form of sampling. So the network will give you some probabilities over some stuff that can happen and you need to pick something out of this now. So this is the the sampling in the output. And if you just take them, the one with the maximum probability. What will happen? That first you will only have always have the same sequence. And it will probably not be interesting. So because the probability of, for example, if you're doing this from. Audio what is very probably is silence so. If you're just taking the the maximum every. Time you just. Have silence, or if you're taking the maximum normal symbolic, you may have. I don't know too much pattern and just repeat over and over. Is that something you want? And yeah, I'm I'm trying to. Yeah, let's see if I can find this one, because I've seen some works on this, but I I never tried them myself, but I can still give. You some pointers. Yeah. So for example this one. UMI will send you. And LinkedIn also.

Speaker 2

Ohh yeah though yeah cause I think somebody sent me stuff in the chat. And that last.

Speaker 1

Yeah, yeah. OK, so I will say out for this. I will say some sampling mechanism, it's what you need. To yeah, I mean this this will need be. Needed in any case, even for people that don't. Care about this? Because otherwise we'll have some just weird. I mean some weird results or some very poor results. Yeah, so people, everyone using these on top. Of the network. And there is not much, but there is some research on. It's a bit more advanced sampling mechanism, but that could give you. Could avoid some of these problems that you're mentioning. I need also to say that this will be also another field, let's say. So people are adding this. Usually they're not paying transformer, they're taking something that is already there and trying to see how well. They can improve. So you will need to decide probably if you work on this. It's harder to work on the transformer. Always person, because otherwise the field is big, but it's it's good to, it's good to know that this stuff exists in any case, yeah.

Speaker 2

Yeah, it's quite a broad topic. You could you could go down a lot of rabbit holes with it. Yeah. So my next question, are there any challenges or specific considerations when dealing with music data in this context, so you kind?

Speaker 1

Yeah, definitely yes.

Speaker 2

Of you kind. Of touched on the. Quantize or the human produced MIDI, but is there more kind of things to consider?

Speaker 1

Yeah. So the the big challenges are the. One of. The music that are very different from from speech data than afterwards. So there is this this long long term structure challenge that that is very big and it's definitely not solved to now. There is the timing challenge which is huge. Because in you don't have this at all in speech and in images, it's not so important. So if you have, I don't know, an image of a car in the background that is a bit more shifted to the right, you will definitely not care. This is still a valid image. But if you have a music and the drum pattern is shifted suddenly of some time not be, let's say correct or even like listenable music anymore. So you need the the system to be. Extremely precise with this timing information, because this is something that we as human we perceive a lot and all these patterns that exist and and this regularity that we have, it does not need to be like very precise regularity as we as I said before if someone is playing it will not be very precise. But it needs to be there in some way that is still not completely understood and not completely studied. And if it's not there, this will just be noise or random noise that you hear around, but it will not give any impression of being music, you know. So these are the 2. Biggest challenge I will say, and they are definitely not. Not at this point. Yeah. And it's something that you cannot steal from from language or images, unfortunately. So that's why also they are they are only there for music.

Speaker 2

Yeah, very unique to to music itself. I am.

Speaker 1

And the other stuff. Sorry for music that is not for for language. Is that we can have this stuff that we have can have multiple nodes at the same time. This is not for language because language is always sequential. So this is something instead that have been studied and people have proposed different solution, but this is still a challenge for music with respect. To this. Language models.

Speaker 2

Yeah, it's something to consider even when you're picking your data. Do you want music that has a note each note, or has chords and?

Speaker 1

Exactly. Yeah. And if you're having an itch note, you will need to start an order if you. Are generating this. If you want to generate 3 nodes, what which one will you generate first? Like? Will you go for the lowest? Will you go for the something in the middle? You know this is. Will you just generate three of them? This is something that that definitely is important because when you are generating the next one, you are sampling from some distribution and then you are resampling for the next one. So depending on the order, you may. Have different results. And this is something that is a problem in in the test generation, because you always know the order of which the words or later are supposed to happen.

Speaker 2

Yeah, that's a good point. When you're dealing with the hyperparameters for the model, are there best practise or guidelines you should follow?

Speaker 1

Yeah, no, unfortunately this is bad. I mean, the solution to this is to try. Everything is really. People cannot try everything so. They kind of go with some. Whatever they copy, maybe another model that they know that is similar and then you but you you will, you will need to do some optimization otherwise you can expect your network performance to not be very good, especially with Transformers you. To you a bit, a bit, a small change can make a huge difference. Sometimes the model don't even train if for example the learning rate is too high. They are particularly sensitive to this hyperparameters, unfortunately, so you need to try to try many of them. Practically speaking, there are some frameworks like I don't know weight and bias. Or tension border. This kind of stuff that that simplify your life when looking for. Is, but usually what you will need to do is like hope to find some good combination then like your network. Some good combination, I mean something this training that you see that the loss. Is decreasing. Then you write your network and then you do optimization on this. With a huge. Set of possibilities.

Speaker 2

Yeah, I was. I was talking to somebody the other day and they were saying that if they were doing grid.

Speaker 1

Yes, well.

Speaker 2

Search that because the model is so big, it could take months to nearly train.

Speaker 1

So that's why this is also not feasible at scale. Let's say this. This will be the scientific way. Now you just do the research channel parameters and you find the best. But this is not feasible. So then you will. You will need to do some choice space and yeah, I mean there, there is no. Suggesting that they can give you because any suggestion will not be very. Scientific again like. So you will find a model that is doing similar stuff. Maybe copy these parameters and and maybe just, you know, do this optimization and. For example, the learning rate of this kind of stuff, that which are very important. But if you have another system that is generating and this is taking this bending dimension for the input, maybe you just use this one. You know, and if you have extra time at the end, you can also try to see. What is better but? Yeah, this is a bit better. I don't like this, but that's how it is. For practical reason, you will need to do some random choice and hope for the best, and let's say.

Speaker 2

Yeah, definitely.

Speaker 1

That there is there any more scientific or? Way of setting this, unfortunately, yeah.

Speaker 2

How do you handle training large scale self attention, self attention transform models for music generation?

Speaker 1

Yeah, you hope to have a big GPU. Do you have some university? Resources that you can use.

Speaker 2

I have my own laptop so.

Speaker 1

This is all, no. Yeah, so this this is problematic because especially for Transformers now, because you cannot really do something. With a super small scale.

Speaker 2

What size TCP do you?

Speaker 1

You can have. A sweet, lovely.

Speaker 2

Roughly use would you say?

Speaker 1

I mean, yeah, it depends what what is available, but maybe at 316? 5% increase of memory inside. But I mean there is there is no specific undercover if you always. The problem is that the the transformer as you probably know the scale with the length of the sequence. If you want to generate very small excerpts like, I don't know next measure for example or stuff like this, you can. You can of course train on your laptop. This will be extremely tedious because you need to. You need to wait a lot and this is. Not very nice. Another solution and you will have your laptop running and then you know this. This is also not ideal sometimes maybe a solution I I don't know what is this. It can be maybe nice to run on this. You know you have this collab for example networks. You know this now? Then you can turn train some. Just stuff. It's also annoying because with the free version you will need to. You will have some time limits. But but I will consider this. Usually I don't know which GPU you have. On your laptop. But you will see so if you if you want to do this just. Aim for some smaller scales, like don't try to to take, you know like millions tokens pieces in your consideration. Maybe just segment the piece and say OK, I'm segmenting the pieces only I don't know 8 measures or something. 16 measures. And then you take this as input tennis. Then it's much easier because if you take the big one, you will have productive dependency. And if you want some batch or or maybe also you can run with very small batch dimension then then it will be slower to 10:00. But you can have longer pieces.

Speaker 2

So are you saying to kind of aim for a snippet of a song rather than hundreds of songs, basically.

Speaker 1

Basically, you can have snippets of songs like inside in the same batch. So there there are these. So if you if you. Care about memory? The problem is that you will. Usually these systems are turning batch, meaning that you don't have. Only one input, but you have a batch mini mini batch of several inputs. The problem is that if you're using, I don't know batch size of. To 128, your system will need 128 more memory. Than a single piece, so you can decrease this, you can decrease it up to the moment where you have one piece only as input. But the advantage? Of doing batches is that it takes also 128 times less to train this, so you use one. 128. Times more memory, but the time is also divided by 120 bit. And if you if you make the. Batch smaller the time will go longer. And if you need to wait on this to do your thesis, this is problematic because we'll take a lot of time now. You can leave it, of course, during the night. And we hope to have some results. But it it is, it is annoying. Of course, yeah.

Speaker 2

So you're saying if the batch is smaller, it will take longer?

Speaker

OK.

Speaker 1

Yeah, because you will need to iterate over more batches. So suppose that you have so you have. Maybe I don't know what other examples. Now set 100 pieces, so if you if you're training with a batch size of 100, basically your system do one iteration. Like one time all the pieces are put into the network by propagated once and that's all. If you are using batch size of 50, you need to do. This stuff twice. In a sequence, so first one and then the other. If this is 25, you need. To do it four times so. The system will take four times more. And use for time less memory. This is the trade off now, so usually people try to to have the best size as big as possible to fit the GPU, because then it will be faster to train if you have a very GPU. This cannot be very high, very big number. Then you will need to wait more time. Yeah. So in theory, it's just about time in, in, in, in practise, people see some differences where they train with a patch size bigger or smaller. But again, it's another one of these things where you never know. Seems to be also a bit random like.

Speaker 2

Yeah, that's actually. The point?

Speaker 1

So maybe maybe consider. I mean, either some just go for some. Yeah, when you need to select what what you want to do, maybe aim smaller because it's also not worth your time to train.

Speaker 2

And then fresh fresh.

Speaker 1

You know.

Speaker

This is not.

Speaker 1

What you you don't need to to have your the next, you know, state of the biggest modelling music ever. You just need to do something else so you. So you learn how to do this. And yeah, show that this. So maybe I'm for something smaller. Me generation of a few measure or this. Kind of stuff. You will still have the same results and you will still. Gain the. Same knowledge particular is practical, more feasible for you without getting mad at retraining this, and then you find that something was wrong and then you need to wait another day and then you know this can be very tedious then.

Speaker 2

UM, that's the point. OK, so getting on to evaluating, could you explain how to evaluate the effectiveness of the model? UM.

Speaker 1

Notice I mean the effectiveness of the model in itself or of the self attention mechanism.

Speaker 2

And I have self touch mechanism, but maybe the model itself like how do you evaluate? In general you've done.

Speaker 1

In general, yeah. Yeah, no, this, this, I cannot tell you. I mean, I don't know. I know there are many. So this is an open question. How can you tell if the generative music is good? And this is an open question and very hard because it's. Yeah. I mean, the best way would be to have some people listen to it and evaluate. Now this is this is the way, let's say, but this is low and then we don't have we we don't want to wait for this or we don't have the resources because this will. Take a lot of. Time. So then people are trying to make some metrics? That can be run faster loss than whatever self similarity metric. To check that the network is doing something similar but not too much because you want something different but also not to repeat exactly the same things. So there are. There are some proposals out there, I actually don't know them because I never had two qualities. I know that there. Are some different one. Most of them work with measure as I told you before. Probably find some of this already implemented some framework, so you can just take them and run them. And that is important to to be aware that they are not. Complete in any case. Now this will give this specific perspective, but for each one of these metrics it is it is possible to find an example of music where this metric will be will be rated very high. That where the music is still super bad if you if you go then and do an example you can always find this country examples. This is not a soft you will need maybe two and of course many of them at the same time and. And be always aware of this and always listen to some examples like most blindly have these numbers and say OK, this is better you know. Because some of these problems can can be there. But yeah, I I cannot. Say specifically because I don't know.

Speaker 2

Yeah, from from what I've looked at, there are a few metrics, but I think most people just get like a survey and they just see how people, how they rate certain aspects of the music.

Speaker 1

Thanks. Yeah, thanks. Yeah, this is this is what you do now. But of course when you're doing some sample at the parameter optimization, you cannot do this. Cannot give to people all the time now. I mean you.

Speaker

In the end.

Speaker 1

Can do it in a scale. Like maybe you have full configurations. You can ask people to wait among the four. But yeah, you cannot do this at the scale that will be necessary so these metrics can be helpful. They can. They can help you be a bit faster on this, but we need to be aware of what are their limits.

Speaker 2

Yeah, that's. I am. Yes. So when assessing the diversity and the originality of the generated music, what specific metrics or criteria do you use to measure these aspects accurately? So I suppose it's it's a similar question.

Speaker 1

Yeah, he said. Yeah, any. This is even harder. Yeah. Again. Yeah, people. People who. Evaluation probably evaluation also be OK now if you're doing a small scale experiments, yeah.

Speaker

Yeah, yeah.

Speaker 1

I don't know what is your colleague. I guess it is to have your thesis done. So I will say as advisor. Then of course you need to speak with your people. But I would say it's OK to if you do it yourself, but you do. It in a good. Way like in a in a blind test where you don't know this, this could also. OK, let's say it's a small selection. I mean, you show that you're able to do these experiments, maybe you do repeated one where you evaluate the same parameter with different pieces multiple times to see if there is correlation because it can also be that you don't find any. This this may be OK, but practically feasible for you to show that you are able to. Do these kind of tests. And you don't need to wait. Then of course, if you if you come up with a very nice system, you can also ask for people to to do. This but. You know, stage is probably not really want to spend the time to ask a lot of people. Analyse their answer and all of this. This will take a lot. To set up these kind of payments, yeah.

Speaker 2

Yeah, that's the kind of struggle. I think that like you're saying, the best way to measure it is by getting people's feedback. But then you go down the route of surveys. And in my company, they said like, they don't mind doing the survey. But then I know the people in the company, so it's maybe not subjective enough for. For the analysis.

Speaker 1

Yeah. Yeah. I mean, if you have people from the company and available, why not? Yeah, but this this is also this is time consuming to to set up these experiments because you need to deal with them in a good way. You cannot have a bug that that invalidate everything. So you need to check this one million time and you. Yeah. And and there are always, I mean this is another field of research somehow because these there are people in psychology for example, that have studied how to properly do these kind of experiments. You know, it's a bit weird and lame also that that for example computer scientist. They still do this, but maybe they do it in a very bad way, you know? So if you want to do it in. A very good way. You will need actually to check this cycle search and see, but this is of course. Not your field again and. Like you know, people have done mental HP's in how to do these evaluations properly with humans. So this we can go. In, but ideally this would be what you will go in if you want to have the perfect one, so you need to be aware of to be necessary and then of course scale it down to your resources. Just aware that this is not the best but accepting some limitation.

Speaker

Because of.

Speaker 1

What is your situation, but not only your your, I mean this is true for all. Or the 15 on the staff. They are all, let's say, not very. Psychology people will not accept them for their conferences, for example, because of the wear there that most of the time. Of course. Maybe. Some of them are very good.

Speaker 2

Hmm, yeah, I suppose it's. It's that kind of thing. It's. Caveats being what the limitations of it is are what? Yeah, but getting people to feedback anyway. So how do you ensure degenerated music aligns with the desired characteristics? So I suppose if you're going down to genres and you know you have jazz, you have classical. And is there. I know it. It's dependent on the data that you're feeding into the model. Is there anything? That the model kind of can do to change it or. UM.

Speaker 1

Here to to ensure that to stay in this in this domain now.

Speaker 2

Yeah. So if you have a specific domain that you want to keep in.

Speaker 1

Yeah. No, I would not know this. Question again like I. Guess no one knows at this point. People are trying to do this controllable generation, so the style could be a parameter when you are trying to control the generation is. But again, I mean finally you you cannot ensure this because the network is sort of free. You can assume this to be true, and if you have for example two styles and you and you input this as a. Input. Then you can condition on this let's say, but you're not. You're never sure about the results, so this is in general for all deep learning stuff. You're never sure about the results, so in this case also, and in this case it's even harder because you how can you check if? This is just. Or not, you know this?

Speaker

OK.

Speaker 1

And other hard questions. So yeah, no, this is fine news for this stuff, unfortunately.

Speaker 2

Sorry, I just realised that we've run over time, but I'll just. I'll quickly go to my last couple of questions. So I suppose ethically, do you see any issues with AI generated music?

Speaker 1

Yeah. So yeah, there could be there could be a lot of issues. So we need to be aware of this. I I mean this is again like one of the other. Places where people have entities future these sort of people studying now. So again we are not the best. People to. To do this, but we need to be aware of these problems and we need to to kind. Of be I mean to to kind of inform ourselves I. Think, but in general I don't think. So as as long as we are aware, I think we we need to set our goals. Now. What? Why are you doing generation now? What is the goal? And I think we need to ask this question. I mean one reason can also be because it's technically interesting. So it. Is fun. For you, for example, to have to be able to train a system that generates some music, this is fun. So this. Can be 1 motivation. Why not? But if you want to do something at least then practically useful for people. We need to ask why? Why would people need automatically generated music? So if your answer is OK because then people that don't have access to I don't know formal music training because of their economical condition, they can still have some fun and generate some music and. And play with this like to create. One thing this is good, I don't think there is any ethical concern in this now. If the if your answer is because I want to like have completely new music that people will listen instead of the music and want, then it's not. I mean, this is ethically problematic, maybe, but not even that is also like. It's not even about that, it's about OK, but why would people prefer this new style of music with respect to what this created by people like? It's people getting something more that people doing something more on music instead of on top of just creating some sequence of notes. These are other questions. Yeah. And and then there is still also then. So this is about the outputs and your systems when they are trained like you need to ask yourself why you want. But I think there are plenty of. Ethically valid application for this kind of. Systems. So it's not about necessarily about substituting the musicians. Now it can be a lot of. Other stuff like helping or or helping people that don't have music training for example or giving suggestion to composers or. Generating for example video games, music. If you want to do real time generations like real time generation is not something that. A musician could do. Because it's real. Time, then this system. Well, the the only way you. Know if you have a. Video game with automatically or a virtual reality system with automatically generate the music that react on what you do. You're not subsisting anyone there, it's just the only way is to have this automatic generation system, so they are all ethical and good application. The other problem is that about the data sets. Now, what are you training for? If you're training on? Data that people have created, they may they may not want to use this data to train another system without you paying them, but of course it's not so easy to to pay them. It's not feasible. Yeah, any clever, for example. So this is another problem and there are big problems, so I don't have a solution. For this, but I think it's it's nice for us to keep this into. To at least know them.

Speaker 2

Yes. Yeah.

Speaker

If you will.

Speaker 1

Be one of these time data sets to train. OK you know that this may be an ethical problem. You're not doing anything still. To help this. But at least you know your reform. So someone in the future will maybe do something about this. Let's see. Yeah.

Speaker 2

Yeah, it's. It's good to keep it in mind. And yeah, and I suppose kind of my final questions is if you were to do this research, is there anything else that you would add that I haven't really asked about that you might consider?

Speaker 1

This what about resources are about about in general, not something that.

Speaker 2

Generating music like just if I was doing this thesis, is there anything that you would consider or is there any question that I haven't asked this?

Speaker 1

I mean, there are a lot of technical questions and that that you haven't asked but but this also was not the time was I'm not a good person. I mean of course you need to to read the paper. This is a bit problematic because there are many people coming out with these topics. Even now I'm, I'm seeing a lot of new. Preprints that will be published in the next month. Some conferences and all of them are different point of view and some are conditioning, some are using different representation and some are using diffusion model, some using transformer token based models. So any you will need. Was to to read a lot about this and also maybe at some point to. Yeah, I mean. Yeah, to to to also try to resume then not to get lost into this. Otherwise we'll never proceed with something. I also think there is a lot of code available nowadays because most people try to put it public. So I mean to start from scratch, just try to reuse stuff that is there and of course then you will need to modify it according to what you need to do specifically. And you need to come up. So either you're doing. Study like, OK, can I do this that this study people have done? This was interesting, especially if you're at the beginning of your career, I don't know. But if you want to do something novel, then you will need to have a kind. Of strong either application or. Idea that other people didn't have or say, OK, I actually think about this use case that these people didn't think and then I optimised my architecture to this use case. So this is not this is not too easy, so you will, I will say you need to have kind of a clear goal in mind like why would you want to generate this music? What what is your? And usage of your system you will need to have this in mind because otherwise you will. Be this kind of system will also be a bit boring to to read and to and to see about because like OK, these people say music well why like what are they doing? Is this interesting? In any case, in any nation except from a technical perspective? And if you just care about the technical perspective, then you will need to have some technical stuff inside that is interesting.

Speaker 2

And yeah, is. There anyone else that you might recommend that might be good to reach out to to talk to about these kind of similar questions?

Speaker 1

People that do, I'm thinking if I know someone that that lead to these generations. That that I think I don't. So because people here are doing some usually more different stuff. So there is a clique of mine that. Is doing. More on this sampling stuff that I mentioned you before. If you are interested I I can give it his name, but it's also I mean it's it's also. Have PhD here. Searching this, I don't know how many people. You. You. Want to speak? With I I don't have. Unfortunately are like let's say someone like a professor or a postdoc to suggest you.

Speaker 2

Hmm. No, that's OK. I just thought I'd I'd ask.

Speaker 1

Yeah, sure, sure. So. Probably yeah. Not not for me. If you if you really need, I can. Of course. Give something, but maybe you have some more specifics from your list of someone. Yeah, I mean, what you can do, of course. You just just find these new papers that you will see there are from 2000. 23 yeah. Are coming out in general. You can always write emails to these people you know then. Then they will of course be very specialised because they just finish our research on. This topic and then they will know maybe more about that search. They will know more about evolution metrics.

Speaker 2

Yes. Yeah.

Speaker 1

And all this kind of stuff.

Speaker 2

Yes. Yeah.

Speaker 1

That I didn't know.

Speaker 2

Yeah. No, that's brilliant. Thank you very much for your time. I'm really sorry. It it ran over.

Speaker 1

Long outside, that was just along in. The beginning, yeah.

Speaker 2

And and yeah, but it was incredibly informative. And yeah, there's lots. There's. It's a huge area to learn from and I think I'm barely scratching the surface. While the sounds of it. But yeah, thank you very much and for responding and everything.

Speaker 1

Yeah, you're welcome and good luck for this work. I hope you can do something nice and that will satisfy you.

Speaker 2

Yes. Perfect. Thanks a million. OK. Bye, bye bye bye.

### Participant 3

Alessandro Ragano.m4a

Transcript

Speaker 1

Yes, no, no problem.

Speaker 2

OK, perfect. Thank you. I'll just I'll record on my phone here. So yeah, I'm not sure if you got a chance to have a look at the the questions, but I've I've put in a another version. So there's they're. They'll follow a slightly different pattern, but they're roughly about the same, but I suppose I just wanted to start off and. And maybe get you to introduce yourself and your kind of background and your kind of knowledge. Yeah. Electric fire.

Speaker

OK.

Speaker 1

Yeah. So I am currently working as a postdoctoral researcher at UCD. My my research topic is on creating quality speech, speech radio quality models that. They're able to measure. Quality of the the applications like video calls or speech codecs or even music streaming. My PhD was on similar topic, but I was mostly focusing on sounder types. Both both speech and music. So I worked with understanding the quality of. Basically, music that is currently preserved. Just like libraries or even. Like broadcasters archives, where they collect many recordings back then, they were analogue so recorded on analogue analogue devices, and now they're digitised. So in all this my my all of. My research is based on mostly machine. And deep learning. Methods rather than traditional signal processing methods. And before that I studied like I have. I studied for my Masters degree computer engineering but. With Special Vision, sound and music computing. So. Yeah. So now it's it's about. Seven years that I've been working on between studying and working in the field of audio signal processing, let's say.

Speaker 2

Uh, no, that's brilliant. That's great. Uh, yeah. So you're well acquainted with with music and the ends and outs of it. Uh, I suppose just to kind of give a brief background to what I'm trying to do, I'm currently doing my masters and so I'm trying to do my thesis on music generation. And so I'll be using transformer models to do that and and kind of focusing on polyphonic symbolic music. So using MIDI files for. Not probably. Preferably like piano music, and that's kind of my focus at the moment, so I suppose I just have a few questions to kind of see. Is there anything I can learn from you from the research that you've done over the years and if there's, yeah, if there's ones that particularly? You know a bit more about. Feel free to. Kind of rant.

Speaker 1

Yeah, just to just a note on that, so I. I never work with music generation. I I work with Transformers. I'm I'm. I'm currently still working on that, but mostly on analysis part. OK, so more mostly for learning feature representations, but I I'm not expecting this. I know, I know some models. I've read, some papers I attended, the many talks at conferences and papers from. Other people but. Are not an expert myself, OK, music generation. Just to be clear, right. So my my expertise is audio quality, which is quality prediction mostly it's an analysis of data rather than generative models. OK.

Speaker 2

Yeah, that's OK. Yeah, I think I think that kind of plays into maybe what I'm working with. So we don't have to go directly into the the questions that we have, but I suppose. From what you're saying that you're looking at more the quality of the music. Quick, so when you're doing your analysis, one of the things that with music generation there seems to be kind of a problem with is is actually measuring the quality of the music that comes out because generally what people do is they just do a survey with people to see, do they like the music or there's different things and it's mostly. Kind of subjective evaluation. But is there kind of evaluations that you do to evaluate the quality of the music that you're working on that could be implemented into this?

Speaker 1

Yeah. So you have. You know many dimensions, so let's let's let's do so. These are like music. Now we have speech. Now, for example for speech quality, majority of people, they would focus on if you can understand what people are saying. OK, this is one of. The most dominant factors of quality and. We can argue that it's less uh, it's still subjective because we are different in terms of the way we perceive sounds, but it's it's less subjective than other factors like pleasantness or you know if you if you want to judge presence, the quality of. This video call. That we're having. You might not. Focus on the fact that you understand everything I'm saying because it's taken for granted. It's like, OK, I understand everything you're saying, but now I I would like to judge if the quality is very. Low. It's very high. With music, so it's more similar here to the pleasantness. Of speech where we need to. Decide which dimension we want to measure the quality of. Music. So one of them could be related to product music production techniques, so that for example you you just prefer this guitar that is more compressed. Less compressed or? Stuff like that. OK, related to the mixing process, the production phase, but another dimension, which is what I worked on during my PhD, was mostly related to the. Noise that is introduced from the bed preservation of vinyl or or or wax cylinders. So very, very old recordings. But all of these factors they will they dominate normally on the music production factors. So what I mean is if I if you listen to two different piece of musics that are produced differently and both of them they have very high background noise like his. Or there's a scratch in the vinyl that you. Know there is always an interruption, that is. Periodical and stuff like. That you would find those disturbances. Is very annoying. And and in that case, that would dominate on the productions because it's like, OK, I I I cannot even enjoy the production of music because there are these problems first which are. More important. Yeah. So this is just to say that yes, there are many dimensions, other dimensions could be even. Emotions or music sometimes can be related to quality, so it also that's very important to ask people the correct question. So for quality we we mean this we don't mean for instance we don't mean if you like the song that that's what they asked ask people not interested if you like the the song. If you don't, that's not what I'm looking for. I was most interested in understanding the quality of degraded by the device and and by the bad bad preservation of these records. So yeah, so there's many dimensions, yes.

Speaker 2

Yeah. So when you're saying like you, you don't focus on whether they like the song but the quality. So like in terms of whether there's like disturbances in the music or whatever. But is there other type of metrics that like the quality because? Part of my research is to, although it's good, when you're generating music to have repetitive music so that you know, OK, the model is picking it up and it's getting familiar with what it's being inputted with, but it could also become very monotonous and very boring. And so to you know, have repetition. Not boring repetition and so am. Is there anything like the music wise that you would have like you would have used metrics for in that kind of set?

Speaker 1

Yeah. So. I would assume that. If you do when you. Say metric, you mean computational metrics like. Two algorithms or.

Speaker 2

Yeah, yeah. Or just, yeah, metrics to kind of evaluate how well, you know, your music has been generated, so.

Speaker 1

It's so nice. One thing you could do is like. Asking like for example in speech synthesis, we use the term naturalness. So what? So usually people we don't ask quality. We ask like. Do you think like something like do you think? Can you judge like? Naturally, speech, so that people automatically they use natural speech as a comparison in their mind, and they try to judge by measuring the difference between synthetic speech and natural speech now. Again, we open new world here because characteristics of natural speech are not only related to the lack of degradations and distortions, but also prosody. Now, for example, if I read, if you let me read something in English and I don't respect the punctuation. Like the period and the. Comma and all that stuff like that. Many people will judge. OK, you know this the naturalness of this is very low. It it doesn't sound natural at all, even if the quality of the speech is is really good in the sense that there is no distortion because some some synthetic speech models, they might introduce some distortions that are a. Bit unnatural sometimes. But in that case, for instance, the process will be very important. So I guess with music you might have similar problems in the sense that what is a a natural piece of music, so for example. I think the culture matters. A lot. So maybe we have to restrict to things with music. I think the problem is similar in the sense that you need to define what is a natural. Piece of music.

Speaker 2

In a sense.

Speaker 1

And maybe so I'm expecting like factors like culture. Would be very important because for someone who is not, that is not used to listen to Western music. Maybe for them piece of Western music is not really what what is not. You know or. Always someone is, I don't know. They they're not used to listen to, I don't know. Electronic instrumental music. And they would expect the colours, you know, the classical structure of top song. So you know what I. Mean I guess. So quality is in general. When it comes to people, my suggestion is always to use references to. Make sure that they are competing against something that is the same for everyone. Because otherwise you don't know what is the process that they used to judge during the survey, because for some people it could be that. Well, to me. I don't know. I just listen to electronic music and to me. This doesn't sound natural. To have the code. So what is important is to keep it. That's one aspect, and then the other aspect I guess is the quality where the reference could be modern music production. This could be used as a sort of, you know, quality reference in terms of. But again, like if the music. Sometimes you know the factors are I. Guess they they. Overlap each other. So you you might have a very. An excellent, excellent recording in terms of quality, but for instance, if the music produced is eternal or you. Know it doesn't respect. Like harmony rules that we have used to. Then it might be very unpleasant for some people and so they will think, OK, this is not not. To me. It is unnatural. Have you? Have you heard? Have you tried? To play with the open AI system, there is a there was a music. Generator that came out. 2020 or something, I guess now we things have improved a lot, but. It was called open AI something. But you can basically upload just a piece of short segment of music and then it produces he. He achieves the complete completing task. I mean, it's not bad, but sometimes to me it produces things that are a bit they have no. Musical taste OK, but I'm a musician, so I don't know if I'm judging because of that or. I I personally didn't like it and amused that was. The name amused net, yeah.

Speaker 2

Yeah, I think mainly because if you're a musician, you can really, it's probably better because you're you can pick out, OK, you know why you don't like it? Like it's the rhythm or it's whatever, where. Somebody else might just say I don't like it and they don't know what's the the problem with it.

Speaker 1

Yeah, exactly. Because in theory there are infinite combinations of no I don't. Know if we try to. Let's say. Represent music as a sort of big dimensional. Data like time and frequency only so. So which pitch you play at? Certain time there are. Potentially many many combinations. That you could do, but some of them. Will not be pleasant, I guess.

Speaker 2

Yeah, I think that's the kind of issue when people do the surveys to ask, you know, you know, is this a good quality piece music when compared to the original or whatever? Is that the music that they generate? They've had a heavy hand in selecting the one that they want to use in the survey. So you know, depending on how many times you're on your transformer, some of them are good, some of them are rubbish you. And then you pick the one you like the best to do.

Speaker

It's actually.

Speaker 2

The survey so it's.

Speaker 1

I understand. Yeah, what you mean? Yeah. Well, yeah, that's.

Speaker 2

Yeah. Yeah, it's hard to know if it's really. If it's it's, yeah. If that test is really testing accurately, but. Yeah, I suppose that's a good a good way to think about it, because like you were saying, if somebody is used to, if they're not used to Western music and a lot of research is on Western music, and then it's just it's, it's not going to mean a lot to you or you haven't got a lot of input in it because it's. Brand new to you so you don't know how to feel about it.

Speaker 1

Absolutely. Yeah, like. We forget sometimes that music is it's very it depends on. Culture a lot like. I mean, you know, there are many studies in that like the music. Innate or you know. Like nurture nature, these kind of things, but. There are some interesting things made on crimes. They clearly show. Like how the emotions we perceive it's because of. The western patterns but. And people don't perceive any emotion after the, you know Elton John song or, you know, if they're not into this kind of. If I think. If their brain is not trained to associate emotions with that in in a sense.

Speaker 2

Yeah, that's actually a good way of looking at it. And I kind of like leads on to one of my like ethical questions is that if I can find it. Yeah, like how to address potential bias or limitations and training data to ensure generation music remains unbiased and respectful of various cultures and musical traditions.

Speaker 1

So, well, yeah.

Speaker 2

Is there a way of? Respecting that and yeah, while you're.

Speaker 1

Yes, one thing is using like. As many jobs as possible, and then it depends on if you. If your goal is. Predict like if your goal is like achieving a task where you have human labels. Then you should be very careful in terms. Of how to label. The music because like some labels, could be just biassed. Even in terms of. You know, think about the jagras like. How do you actually define the jar like do? You take it. From sources like Wikipedia, the Oscar Fashional musicians or journalists, I don't know, like anybody can have a different opinion on the Java I guess. So I I. Now I don't know about this generation because I never did experiment myself, so I cannot tell you things on. That part. I can point you to paper I. Was a professor at Queen Mary. You wrote this paper. Oh yeah, OK. So. I think it's an interesting read because and so in the music information retrieval community there was a paper that was used a lot called the the sound. Sam stands for the surname. Of the author, which was at Sanitatis. And then he was, you know, one of the first things done with the air through machine learning using machine learning. So we did, we didn't have all the knowledge we have now about. Bias and stuff. Like that but. Yeah, there's a theatre that. Yeah, it's called the. Yeah, the JIT, some data set, its contents, its faults, their effects on evaluation and its future use, which showcase like all of the problems that are in this data set, I think it can be a good reading to understand, OK, what is it that can be wrong? And building a data set or music and they talk about all of these problems, like repetitions or music or using like for instance one genre like music, always the same art. Stuff like that because. Ultimately, the goal is removing the as much as you can come hundred factors, so you would like that the model predicts certain things not because of other factors, but exactly because of the. Variational factors that. Are the the calls behind that? Value of the label. That's yeah, but of course it's very complicated to track in the in the neural networks and to understand this. Yeah, I don't think we're able to write down the name of the paper because, sorry, I'm. On the computer. Otherwise I have. To open it from the iPad and.

Speaker 2

If you could you send the link to me on LinkedIn cause then I cause if I'll I'll lose it on this chat.

Speaker 1

Oh yes.

Speaker 2

It happened to you last time.

Speaker 1

Yeah, that's one thing about music generation, I guess is the diversity now. I'm just thinking out loud, OK, because again, I I didn't. I I never did expand myself so, but I'm expecting that definitely the variation in music is very important because otherwise. You know, you might have. Always the same output or. And also I guess do you? Do you work with condition condition based music generation like do you condition it with something or?

Speaker 2

Yeah, I think. Ohh, I've I'm still gathering up where I'm. Doing it but. Yeah, I think it's kind of like to make sure like say, like your the notes are being hit. So like, say you have it in a certain range and. I can't think of it now, but yeah, there's there's conditions that I've I've lost my. Train of mind. UM. Yeah. And I think I think Ethan like. Oh, it's it's not my head. I'm. I'm going to start off for now. But yeah. Yeah, there is conditioning in it to kind of prevent. Mundane reputation like temperature and that.

Speaker 1

I guess so. That that's it's. Important. Yeah, to to have radiation like, now I I didn't read the the the recent papers visit generations Transformers. I remember the first papers that came out in 2019. Something that and they were using like classical music datasets. With, you know, media recordings, so basically three or four voices, the most quite simple something maybe can be useful for you is. Need some experiments in terms of. Creating the so. Easy model that. Is quite works quite well. In speech analysis and not in the analysis, is the way to work. It's a model that was developed by Facebook and.

Speaker

He he what?

Speaker 1

He can do. I don't know if you're familiar with these things called general purpose representation models.

Speaker 2

General purpose I don't.

Speaker 1

Yeah, like so, these are models that they. So they through. A certain task. They transform the input features so that runs or the waveform into some features. That they're able to adapt well to several different downstream tasks. Yeah. So, I mean, what I mean is like, imagine you want to do. Music, music, genre recognition and picture recognition. OK, these are two different tasks, OK, normally what you can do is you take features like Mel spectrograms, you train your CNN to do the first task, then you train another CNN with different labels, not to do the second task. Now one trend. Is to create general purpose models, which is. You take a very large. Data set and a very large model, and then you ask the model to. To basically learn a. Task that is none of these task we labels OK, it's just a task that is about recognising the structure of the music itself. Now I can give you some examples. Later on but. So through this task you. You train this model by using a. Large, very large labelled data set. And then you basically take the features from this pre chain model, you adapt it to to several several different tasks. Now if your. Features are good enough, they're able to adapt to all the tasks because they understood what the what the music is or what speech. Is in terms of structure. And I worked so with the. I I evaluated like speech model. If was able to work with music this wave to like model I told you before. And what I did is I just pre trained it with classical music only. Even if I. Use every classical music like. Then I fine tune the model on two tasks page recognition and instrument recognition. It it was really good at recognising like instruments that they were never seen in the. Classical music data set. OK, so sometimes. Models is able to adapt to out of the main distributions. Quite well. If it's flexible enough in the in the representation. And it is just very. So that I did experiments myself. So with the classical music and. He was able to adapt so. I would I. Would focus on minimising, not minimising the like. Understanding what is it that you're trying to achieve in your thesis in the sense that without, you know, without thinking that you have to solve everything like just. If this little. Task can be done with classical musically. Then you just use that. It's more simple and the data set is available and. It might produce good results even if you're interested in other type of data, yeah. That was my point. Yeah. Sorry to be bit said many things.

Speaker 2

Yeah, I think no. No, no, I think that's good. Cause I think like you're saying, once you start going into bigger data sets with lots of different music in different cultures and it's not classical music, it just adds complexity complexity to the whole. To the whole process and it can be harder to do. It seems like a very unknown kind of. There's still a lot of research going on to kind of get a bit more, because I think everyone's just kind of trying different things and seeing if it will work.

Speaker 1

Yeah, totally.

Speaker 2

Yeah. No, that's great. I suppose like leading into kind of a more ethics question again is like considering the potential of AI generated music in various contexts. What ethical consideration should be considered to ensure responsible and and respectful usage of the technology so I know you haven't done music. Generation Bush is a the kind of things to be aware of with ethically, with AI generated music, do you think?

Speaker 1

Oh yeah, it's a. It's an interesting question I. Well, I guess. I guess, responsible. Respectful. They mean like, you know, they they depend on which context we're talking about. Like, again. It's. Yeah, I would. Say it's about culture like. Maybe some music can be considered like as offending for some people lyrics no factor that's important. Some people you know, some users can be considered vulgar or because of lyrics or. So all of this. Here is that is important and then there is. The problem of. Copyright. I guess you know that's what you mean as well, yeah. So you know music before. They are musicians have sued themselves like each other.

Speaker

For a long time.

Speaker 1

So this it tells us that the problem is. It's more the way. We do things that sometimes it's about. We get we. Get inspired, you know.

Speaker 2

Like, yeah.

Speaker 1

That's that's the yeah. So I presume that. Synthetic. Yeah. Generated music. You can definitely repeat things that are in the training data. And definitely it's important to have music that is not protected by copyright. If you want to. Avoid these risks. But then the question is. Like how much music we have out there that. Doesn't have copyright. There are some. Efforts made by the Mir community then. I don't know if you're familiar with the free music. Archive data set.

Speaker 2

I'm not familiar. With it, but I'm aware that if something's is a certain period of time, or like it's very old music, you don't have to get copyright or something like that.

Speaker 1

Yeah, there's just some other you can also find. Modern music that doesn't have copyright as well. But I I I work with this data for example. I listen to some samples. Sometimes you can perceive that it can be different from professionally recorded music. There are many noisy recordings. For example in the Free Music Archive data set. So these are these are can be problems like in a sense that because you have you know professional music then normally it's protected by copyright. You could use that to train your models because you know ultimately you're not sharing your own, you're not sharing with the data. So you're protected you as a. Researcher. But then the model will probably output something that will. Easily, you know, resemble something that is protected by copyright. To be fair, I don't know how to solve this problem, it's I think it's a huge problem that that people have like in in the. Music industry for sure. On our side, what we can do for our own, you know, small projects like it's definitely using music that is not protected by copyright, especially if you want to share your model on GitHub and. Because you never. You never know. Like you never know. It's responsible in this case, so.

Speaker 2

Yeah, play the safe side until it's figured out.

Speaker 1

Yeah, yeah, that's that's important, I guess. Yeah, let me think if there's anything else. That's what I can think of in this moment. Yeah, but no.

Speaker 2

You mentioned earlier that you have worked with transformer models yourself, so there's various different ones. There's excel Transformers and there's vanilla and what have you, but. UM, is there kind of like, how do you kind of like measure how effective that transformer has been? Is there specific metrics that you use or like have you and even like more specifically measuring how effective the self attention is in that transformer, is there kind of metrics that you go for? In that way.

Speaker 1

So usually what what they did is for instance comparing models. So the classical structure is you have you segment your input signal, then you learn local features, for instance with CNN and then you. You take all of this. The output of the CNN which are local features and you you use them as input of the time dependency model. So a model that learns how to keep learns what is the dependency of each local feature with the other local features in the signal in the long sequence. OK. One thing that easily is to do is like OK, I fix my CNN backbone. To extract local. Features and I just change the model. For example, first I use LSTM, then I I change the LCM with. The Self Attention network. And I see. If there is an improvement in the. Performance. That's the most simple thing that I will do, yeah. And very often it turned out in my case that it's useful to use self attention network. But it is also important to understand what is the input features that you give to the self attention net. So in the audio community, some people have tried to use Transformers without CNN as a as a you know, the step before the transfer. So without giving the CNN output transformer but but giving directly for instance. And it's very, very hard to train a model an audio model without some local features that are able to learn what happens in 25 milliseconds of sound. So there is a paper on that. They have some good results, but they need. To rely on pre training models. So they use a transformer that. Has been pre trained first. But if you want to train for scratch model that is not spectrums or waveform or whatever transformer output. It's very, very hard. Like it's it's still not. It's not. It's not possible to do something like that and getting the decent results the same. So that's one thing. Yeah, that we'll do is using CNN and then. Understanding which is the best way to model the time dependency. Also you could also compare like against other. You know you could use LSTM Self attention network. The transform you know. You could also. For instance, use like just. A simple average to have a certain baseline. You average the output. Of the segments of the CNN. Just to see. If modelling the time dependency is something important for your task, like for quality for quality is very important because quality usually degradations are time varying so they change the statistical properties change over. The time, OK. So if. You have one minute or. Each the degradation statistical properties that you have are around the 2nd 10 and then around 2nd 40. They wouldn't be the. Same so when. You have a time dependency model, you're. Able to learn this. That's that's very important. Yeah. But I guess other tasks, maybe they're not as important as they seem. For sure, music generation is important because what you generate at a certain time depends on what you did before. So there is, yeah.

Speaker 2

Yes. Yeah.

Speaker 1

And one thing that I noticed from attending talks and conferences is with music generation. People use what is called the relative self attention.

Speaker 2

Yes. Yeah.

Speaker 1

Yeah, I guess you you know that, which apparently. Works much better. With music compared to speech or NLP tasks. So yeah, you had another question like why does, why does? It work or something, no?

Speaker 2

Ohh sorry, it was just like to evaluate the effectiveness of the self tension mechanism that which you answered. And then just kind of in general how well like how to what kind of metrics you would use to evaluate how well the transformer as a whole? The works.

Speaker 1

Well, to me I'm not aware of any metric that you can apply to, let's say, interpret the features. That there's some loans I'm not aware of any anything like that. What you can rely on is definitely understand measuring the performance of your tasks. So understanding like the output if. If what what you know? What? What? You're getting like something that? Makes sense or not? Now I guess because you. Are working with music generation so. But your problem is how do you measure the performance of that right or.

Speaker 2

Yeah, I think they kind of like they do clustering to see. How close it is to the centre point, but also how dispersed out it is like it's close but still. So that so that it's similar to the input, but it's varied and it's not like the OneNote all the time. So you're kind of that's kind of what I've heard so far. So apart from the kind of surveying and getting people's opinion.

Speaker 1

I guess I guess people. Opinion is I mean like with quality is always. So it's better to have the. President, sort of, you know, the preference that. At least you know for sure that. If you made a good survey. Then you know for sure. But the.

Speaker 2

So when you're doing surveys, is there kind of like a best kind of way that you would go about it because. I think like I was saying before, like you can pick a piece of music that you like the best, that of the generated music and then compare it to maybe the original piece. And do your survey that way. And like people I've spoken to before, they kind of did their survey with people that they knew, like just hang out to friends and things like that, which I think you're supposed to avoid. So yeah, it's kind of like. To get a good uptake on people actually doing the survey and then yeah.

Speaker 1

Yeah. So. First thing is. What I would recommend is creating the test where you the most important thing is understanding what is the question you're asking and you need to make sure that what you're asking and what you are getting from people is what I mean is like you need to make sure that people answer based on what you're asking. Be some something else. One suggestion I would. Have is to run pilot tests first.

Speaker 2

OK.

Speaker 1

With someone that you know, a friend or whatever, or when you know someone is, if it's an expert in this field, it's even better. You don't tell them like you don't bias them, you just ask them to do the tests, and then you talk with them after you explain what you did in your in your models. By doing the. Pilot test is try to evaluate all of the things that you are not sure about. Yeah, one of one of the. Things that could be the question. You're asking even. Sometimes be careful with the reasons that people tend to judge things relatively to what they have. What I mean is, so if in your life you always had the excellent food from the best chef in the world. Then one day you eat something that is pretty bad, you know, student food or whatever you'll be like. OK, yeah. You'll be very you. You wouldn't like it, but if you don't like you, you've been eating student food. Maybe it's like. OK, for you, it's. Acceptable. So things are very. Relative in the listening test, so make. Sure that you introduce a lot of variation. In the stimuli. So that the people can, can they we use this term elicited? No. Like, they illicit illicit. Yeah. Like you can can you say in English like.

Speaker 2

OK.

Speaker 1

Evoking. Yeah. Yeah. Yeah. Like, they they can appreciate that they appreciate what they can perceive, that there are differences in the in the evaluation of the stimuli.

Speaker 2

Oh yes.

Speaker 1

Because if you don't introduce. If you if you introduce. Very small differences. Many people can easily judge. There's no difference here or OK, so make sure that yeah, you you have a variation.

Speaker 2

Bit like an eye test. They're always like one or two and they all look. The same.

Speaker 1

Yeah, exactly. Yeah. Yeah, that's that's important. And I would also make sure that in the case of music generation, to understand, if you're looking for expert people or non expert people, this can have an effect on the on the results because. Expert people, which can be. Either researchers in music in generative music or. Mir in general. Or physicians? They might be very picky and giving you a judgement that is, you know like. It's sometimes you. Often it works better because they're able to. Capture nuances that the naive person naive listener wouldn't be. Able to do like. So if you have, if you know people or musician, you know what can they see? It is better to to. Get data from them. And to report that. That these were experts. Yeah, because otherwise I. Don't know like. You know if. Someone is a casual listener and they never paid any instrument in their life. They never. Did anything like that? UM. I guess, yeah, the results will not be effective this this, I mean it's done like with quality for example. It's very important to make sure that because. Casual listeners are not able to differentiate some quality characteristics between different stimuli, and I think it's the same for generated music. It could be the same.

Speaker 2

Yeah, I think you're right. That's kind of. I was thinking the same. Uh, when doing the survey to kind of get more musicians themselves to look at it because. Like you're saying, like they're going to notice, uh, you know whether it's a normal cadence or whether that's like a very long break in the music. And it just sounds ridiculous to have like a full. Stop there. So yeah, or.

Speaker 1

Other things they have as well is all the suggestions is be careful of fatigue during the test because listening to music it can be tiring after a bit, especially when you change the music all the time. Because you know you have to switch from. One song to another and. Well, if it's needy, maybe it's less tiring, I guess if it's a real music, it can be more complex as a task. So just in the pilot test, for instance, you could ask at the end if people experienced some fatigue during the test or not. Just to just to have an idea of like OK timing as well is important because. So you know. During the pilot test, if you if you have a way, I understand that sometimes you don't, but if you have a way to measure how long it takes for every participant to. Finish the task. Then you have you have an average and based on that you can decide in your physical test to remove some tracks, maybe or to add more. So you you you should. Realise like if your test is not long enough and has duration and also it's not very. So that's important, because otherwise people might give up or there are many studies showing that when the test is too much tiring and. Too long the. Performance of people dropped significantly in the last stimuli that they. Have to judge. So be careful with that. As well. And then I guess another thing is. If you're planning to run the survey, your line or online OK, because you know, if you, if you ask people to come to the office or whatever in the lab. In that case you. Make sure that. For example, you use the same headphones you everybody, so anything can buy us the listening test, OK, anything. Even the interface, the computer use anything. OK, so when you you are in the lab, you make sure that everybody's using the same equipment. You can see what they do. You can answer questions if there are any doubt when they are answering the task, maybe some people don't understand well what to do. Some people don't understand where the question always seems. So when you are there physically present with them, you can clarify any possible doubt and. So when you are not with them. In this case, like, I mean, just consider that some people might be affected by these. Things now I think for your your test is fine like just. Just consider that OK that there might be variability introduced by these factors and.

Speaker 2

OK.

Speaker 1

Ohh, one thing you could do, you could ask things like if they they have healing ability like you just ask them if if you know if they. Have an issue with that or. Because like for example all the people. They, you know, they don't perceive anymore higher frequency, which is. It's a good. Part in a in a piece of. Music and so. This can affect judgement as well and. Maybe you could take some. Demographic data like about age, you could do some data exploration to understand if there is any difference with age or anything. Like that or. Now I mean, look, you know we we don't like when doing this test. It's not that we can control everything. So obviously if a company they they they have to release a product they will you know they will make a survey that they measure everything as possible. But sometimes you you you can make assumptions and you know, just ignore some of the problems. But you are aware of it. You are aware that's fine. Yeah, because even you know the devices. They can affect. Like the perception that you you might have, you know. For instance 1. Thing you could you could do is. Asking them, I used to do these things like, you know, tell people you must use headphones, you know? So like, once they did an experiment on Amazon tour, where I I came to light, I said I have a mechanism to understand. If you don't use headphones it. It's true, but. I think it was an incentive like because yeah, because, you know, people are lazy or you know. It's better if you tell them look.

Speaker 2

You have to.

Speaker 1

I can. Yeah. You know, you're so so they pay attention by the way. Like, that's another thing that you have to be aware of. It is. If you have a test that is long, there are some mechanisms that you can use to understand this. Participants are paying attention to the. And one of them is hidden messages. So that, for example, you know you have first the tests start with the piece of music, then another piece of music and song again, but the 4th. 4th track is not the song anymore, it's something that starts with two seconds of a song and then there. Is a message saying. Click here to confirm your attention. Please something like that? And then if they click there you you know that, OK, they're paying attention. They're not just answering, answering randomly. OK, this this can be. Yeah. Something some people use. If you want to know.

Speaker

What about that?

Speaker 1

Like let me know I can I can. I can send you some some. Material that where some. Researchers have used these techniques and. Also in terms of. Yeah. One thing you could do is. If you don't want to use anything like that, you can also do that once you have the results, you can analyse participants out there and for example, you can exclude participants who didn't introduce too much variation in the.

Speaker 2

OK.

Speaker 1

I don't know. If you have 80 tests and everybody. Answers always say.

Speaker 2

OK.

Speaker 1

There is one participant. It's always a definitely. That person is not paying attention to the task. So OK, you could. Do these these kind of things OK? When you don't have too much variation. Clearly, the stimuli that you're using introduce variation in the judgement. There is something wrong with that participant, especially if other participants they're not behaving the same. So some people, they do that they exclude. Are reliable, you know are reliable participants from, from from.

Speaker

Right.

Speaker 1

OK, now I'm telling you many things that just. Yeah, make sure. Like there are many, many, many things that you can. Look for to make sure that. There is also you have make sense and.

Speaker 2

Yeah, I think they some of them have been brilliant, just even like making sure they're using their headphones. But even just having that short clip, that's it's not even that big a deal to put in, but it would affect, you know, it's a good way. Of figuring out. These are each participants.

Speaker 1

Because I I tell you from. My experience like. One for one large Amazon to leasing test and many people didn't get this answer correctly, so they were just playing for, you know, they're just cheating in a sense. Yeah, so many of them, like, they found a lot of them. And and I didn't pay them because that's the rule on Amazon turkeys you have. Conditions that if you don't respect them, you will not get the money even. If you do the test. OK, I lower than at the beginning, so.

Speaker 2

Yeah, that's actually that's very good. Yeah, I suppose. All right. So how you got participants is that you said this, they would get a certain amount of money if they completed the survey kind of thing.

Speaker 1

You could, yeah. You could use Amazon Torque for yeah you need. To create a task. It's very simple if you know a bit of HTML. Or there are templates you know sometimes. Which they can be helpful.

Speaker 2

Yeah, because I was going to try, like, SurveyMonkey or one of those kind of ones. I hadn't.

Speaker 1

Nice. So my supervisor group. They they make, they made an application called Go listen.

Speaker 2

OK.

Speaker 1

And if you if you type. Google go lists and UCD. You will find. It's an application that it's very easy to make a listening test with it very, very immediate and you you just get. A link that you can share with people. And it's very, very simple to me to. So my experience is the most simple way I've ever seen for making this interest because it's with an interface where you just upload tracks and. Now, of course there are some types of tests like AB something called ACR Musha test or I don't know if the test that you want to do will fit this templates, but take a look maybe.

Speaker 2

Maybe that's that's actually really helpful. Because like that SurveyMonkey. You can you can put in questions, but whether you can put in your clip and kind of do things and do. It well. Is another thing.

Speaker 1

Yeah, I see.

Speaker

Yeah. No, it's.

Speaker 1

It says, yeah, I I see that some people have used it but. Yeah, it can be limiting against.

Speaker 2

Yeah. No, that is extremely helpful. Yeah, I was trying to.

Speaker 1

Looking you have some. Questions here on the self attention.

Speaker 2

Yeah, I don't know. Is there any that's you kind of stick out to this you feel comfortable answering. I know we're kind of coming close to the end of the time.

Speaker

Oh shoot.

Speaker 2

But I didn't want to kind of delve. Into it. If you know music generation isn't your. Specialty. I don't want to put you on the spot.

Speaker 1

Yeah, yeah, yeah. No, it is, it is not even if, like, so normally yes. Like I see that LSTM I remain the. The main problem they have is you. You know that. You first of. All the classical renown there is the the problem of Ironman is the problem of the vanishing gradient or exploding gradient, yeah. That's that's the. Biggest problem, OK we LSTM you solve that? But still. The problem is. You you know, you have the representation token one, then Token 2 depends on token 1. Token 3 depends on token 2. That in turns depends token 1. So you're accumulating like that. He said the the, the the thing with the transformer is exactly the capacity to learn the dependencies between each tokens. You learn that with the. Task and. So LSTM has no such thing where you you don't have like you know, it's not at LSTM is able to say or I need to pay attention for the token #10 I need to pay attention to the token #3. He doesn't do that at all, like talking 10 takes information from everything accumulated up to token 9 plus, plus another increases in translation translation tasks you will take also the original token feature and then you want to translate. Into the new language and. Yeah, that's that's the. I guess is the most important thing. So you know to have this mechanism of key queries values where you learn these three matrices and. You have, you know, the the queries, the token you're looking at, then you have. A bunch of keys, which are all. Of the other tokens and then you. You learn the weights. By doing the the dot product between the two and then your sum the. Weights to the. Values tokens which are again all of the all of the tokens of the set. But this time, thanks to the weights that you learn with the product, you can carry the query and keys. You will you will get the importance of. Each values with respect to. That particular token that you're looking at. It's easier to explain with maths and rewards but.

Speaker 2

Yeah, yeah. UM. I think the rest of them are kind of. Very much. Generative focus. And have you ever worked with MIDI files yourself?

Speaker 1

Sorry we did it.

Speaker 2

MIDI yeah.

Speaker 1

For I did here, but I did when I created that model, they waved back for music just just to fine tune the model to classify instruments and pitch.

Speaker 2

And was there like specific maybe files that you were after or was any kind of pre processing that you were kind of focused on when you were doing that?

Speaker 1

Uh, no, I I've used the the data set called N synth. She's a yeah, it's a media data set. Yes, it's synthetic and it's set and you. You basically have labels about velocity, the MIDI number, and all of these properties. Yeah, yeah. No, I I because I I've used it only for. Yeah, for fine tuning directly but. Because when you. Say me, do you mean do you mean Wi-Fi files recorded with me? The sound is that what you mean or?

Speaker 2

Media as in like where it just stores the information rather rather than the audio like there's no audio to it. It's just like I know she's played at this time.

Speaker 1

Yes, so directly using. Directly the values like.

Speaker 2

Yeah, it's English.

Speaker 1

OK, so the. Velocity value the media value and the the the time value timing.

Speaker 2

Yeah. Yeah. So, like, you can't audibly hear MIDI. You just it's.

Speaker 1

No, no, no, no. I never worked with that. What I what I worked is. Wave files generated with MIDI instruments, but my input features there were wave files.

Speaker 2

OK. OK. And yeah, I think that kind of. I do. I I. I think the rest of them. Are kind of more. Generative music questions, which is fine, but I think kind of your expertise and. The quality of music and even doing the survey, that's a huge help because I think I've talked to a few people and we haven't got that deep into those questions. So I think that's that's massive. And but yeah, and and I don't want to hold you up because. I I know we've we've gone on for over an hour so.

Speaker 1

That's no problem.

Speaker 2

Yeah. So I just wanted to say thank you very much for for taking the time out to talk to me and for answering all my silly questions.

Speaker 1

I hope, yeah. I hope it was useful.

Speaker 2

It was definitely.

Speaker 1

If you if you need any help for this is or whatever, just let me know for the test or you know whatever. Whatever like you know the things that we discussed it. If anything is not clear or. OK, so the you will not. Bother me, so don't worry.

Speaker 2

That's great. And thanks a million for that and yeah, I'll see how I get on with this and see if I have any questions. I'll, I'll let you know. But yeah, thanks a million.

Speaker

OK.

Speaker 2

Enjoy the rest of your day.

Speaker

You too.

Speaker 2

Seeing it bye bye.

Speaker 1

Right.

### Participant 4

Vincenzo Madaghiele.m4a

Transcript

Speaker 2

You thanks so much for for replying to my message and for agreeing for the the call. It was really helpful. I just want to double check. I know I've asked. You're ready, but you're OK with me recording the the meeting?

Speaker 1

Yes, yes, that's all right.

Speaker 2

Yeah, that's grand. I was just wondering, did you get a chance to have a look at the questions? I know I think I sent to you another. Set of them.

Speaker 1

Yeah, I didn't. I didn't manage to download them. Should I download them now?

Speaker 2

You can, if you want. I'll. I'll be going through them anyway. But if you want them in front of you, you can. You can see them then.

Speaker 1

Yeah. OK. I'm going to download them now. OK. Yes, I see them.

Speaker 2

Yeah, that's grand. Great. So I suppose we'll start at the top. And do you want to just, like, introduce yourself and your kind of background and? And kind of what you've kind of studied and what you're. Working on at the moment.

Speaker 1

Yes, yes, yeah, I'm. I'm I. I studied engineering software, cinema engineering and media engineering and data science. And I worked as a researcher in sound computing. I'm also a practising musician and sound artist. Yeah, and yeah, at the moment I'm working as a machine learning engineer at Stockholm University and I will start actually PhD in sound and music computing and also university in November.

Speaker 2

OK, very cool. That's great. That's perfect. So I suppose we'll get stuck into the questions. So kind of just to start off easy enough. What? Uh, so in my research, I'm going to be using Transformers. Have you worked with Transformers?

Speaker 1

Yes, I I'm. I'm assuming that you contacted me because you have you seen my project, my paper on music generation is it? Right.

Speaker 2

Yeah. So yeah, just going through, I've contacted a few people. So have everyone muddled in. My head but. Yeah. So I suppose to start off kind of a a basic question, what makes Transformers suitable approach for music generation compared with other methods like LSTM and RNN?

Speaker 1

Yeah. So, I mean, if you see my paper I have on machine learning English generation, I am comparing transformer based approaches and and current neural networks approaches. They are good for different things. I think if you so like they're both suitable little different things. I think from my and that's the other interesting thing is that like the the understanding is pretty much intuitive. So the. Just from my experience, I talk a little bit about that in the paper the they're good for different things that it's difficult to put into words because the transformer understands certain things because it's it's optimised for language. And arenas, LSTM, they are used. They are more Chennai so. They're kind of like. Transformers, in my experience I used it because I wanted to put music theory in the model, so my model has embeddings for pitch, for duration, for bit, for. And in that way I can describe music as chords, for example because my project was about jazz music. So I needed like a jazz structure. So I was using Transformers to to describe codes. And that I can do in Transformers because of the self attention mechanism, so it can with the embedding it can learn what for example pitch is the concept of pitch and then if you describe the colour as being like as being. An aggregation of four pitch using the same embedding. Then the model knows that the cord is like a a set a set of pitch and it has these properties and it understands how pitch relates to each other. I think the networks are like LSD and these kind of modified models. They lose a little bit in in in the how much into how much you can call the theory directly in the architecture. OK, so I don't know if I'm clear in.

Speaker 2

What was the last? Thing you said is that the cover theory.

Speaker 1

Yeah, in my. Case I was theory, meaning what is the pitch? What is a? Duration so ultimate. What is that I call encoding, music, music theory in the code because any states like Western music theory from jazz in my case. But most people this is Transformers do that also they use Western music theory to encode that in. Their mothers.

Speaker 2

Yes. And so how did you go about that? Is it, is it mostly in the preparation of the data that you can encode the music theory into your uh transformer?

Speaker 1

Well I am. It was both in the data preparation and in the model architecture. So if you look at my architecture of my model. Take I was using. Symbolic music data set. Mm-hmm. So notes. Represented, I think I was using music XML format, yes. So the music theory is like I'll take every note and I say like and I extract features. So for every node I extract the pitch. In MIDI format, the duration of the Note. So how long it lasts? Yeah, and the the beat of the bar on which the note is. And this is so for example, if you notice on the first bit it has a different in music theory that I was referring in, the music theory was referring it has a different function in the method usually.

Speaker 2

So if it's the first note in the bar, it's that's important to know, like if you're playing C at the beginning or C in the. Middle it's a. Different. This is what's happening, OK?

Speaker 1

Yes, yes, depending on enjoys the case depending on the. Court, that is. That is played. Usually you can find that the significant amounts are in the beginning of the bar or have played in rhythmically significant positions, let's say. And and that's for another job system.

Speaker 2

So if it isn't like you're saying you know, there is significant positions. So apart from it being played at the start, is there other positions that would be significant as well?

Speaker 1

Well, I would say that like all positions are significant, but they have different functions. So for example, you know if you're familiar. With the concept of. A passing note. Yeah. So if a note is in the like, let's say the if it's a cell phone maybe. You encode the bits in there, so I was using different bits. For example, it could be like 01/2 or three basically so and if you notice on the one on the first bit on the third bit this. As is usually interpreted as like strong or not, it's usually like a. Because, but that's also. That's complicated, because it also depends on what the drama is playing. So this is like all the small things that you cannot really like that the model has to learn from the data. So you cannot. I mean you can so this this way of describing things. This is help to the model. And that I was doing because because. I have very small data like kind of small data set. So it's I I thought. Of like being. Like very explicit and helping the model, giving it more information from extract that can be extracted from from the the music score. OK, yeah. Now I don't know. If it's clear.

Speaker 2

Yeah, no, that is clear. And so when you were saying that you were putting attention on the 1st and we'll say third note roughly. So was there, was there a way that you tuned the attention to be kind of focused on those notes or?

Speaker 1

No, no, I I don't say that these are more important than the others. I just say that for each note in the sequence, if you imagine a sequence of of of characters, right?

Speaker 2

Yes. Yeah.

Speaker 1

You have and like so each each element in this sequence is a node. It has a pitch, it has a direction, and each play in a specific point of the bar is not. So what I tell the model is is just. I don't say the first bit is more important than the second one. I just say this note is played on the first beat and this. Note is played on the second bit.

Speaker 2

Yes, OK.

Speaker 1

And the function of this. Note in the melody, it has to learn from itself. That may be notes that are played in the second bid. Are different than the ones they played on the first bid or not. The style of music and on the data basically that it has, but the the what I call encoding music theory.

Speaker 2

OK, OK.

Speaker 1

In the fact that the family just give it the information, I say that bit is a concept. And bit is important and that's already helpful for the model because then it can understand like yeah bit. Has or not. It can also not be important because of the structure blah blah. And that's also kind of in my. Paper I compare like how it performs. With and without this information. And the general the music is different, but it's very difficult to say if it's. Better or not? Yeah, it's quite subjective. Yeah. Basic. Yeah, that's that's kind of. Yeah, exactly. You can, you may. Like it or. Not it can be like, but it's also weird. It's just I think it's a style.

Speaker 2

Yeah, yeah.

Speaker 1

It's just different style and you can like basically, yeah.

Speaker 2

Yeah, exactly. Yeah, that's brilliant. So I suppose my my most of my questions are quite general, but I suppose from your own research you'd be. I'd say you'll you'll be able to answer them, but so for my next question, why is self attention critical for music generation? What unique aspects of music make this? Mechanism important.

Speaker 1

They need the attention module in the transformer model, right? Yeah, because you you want to work or you are interested in one. Yeah. So encode the modules with self attentions, am I?

Speaker 2

Yes. Yeah. Uh, encoder. Yeah, yeah, yes, yeah.

Speaker 1

Is is interesting because it can it kind of deals with the. So it has this matrix right of and it's basically sort of like a fancy. From my understanding. I'm not extremely like extremely focused on the mathematics and like how it affects, but it's basically a fancy self correlation correlation metrics. This importation model, it's a complicated way of saying that one category of objects. Which is can be a pitch or duration is related to another pitch. For example one pitch can be more related to the other. There's, I mean, this is my understanding of how I can like it sort of interprets self attention. That's interesting because it's kind of like allows to. Yeah, that's like. To like define. Define like sort of long term correlations I guess. So that can be a good thing sort of long term correlation that can be a good thing. But also it's like. I don't know. Like the. Yeah, it really depends on how complex the data is because like, in my experience that also can lead to, for example, boring music because the notes are everything is very generalised.

Speaker 2

And so do you think having having self attention, although it kind of gets the essence of the music that it's been trained on that it might be too boring or too repetitive of what's being trained on? Is that what you.

Speaker 1

OK.

Speaker 2

Mean or?

Speaker 1

So self attention is basically the probability that after one node there. Is the other node. So that's the the matrix contains the probability that the next note from like if you start from this note most of the probability that the next one is CABABCD and so on, and then you choose the most probable 1. And those probabilities are learned. From the data. So basically it computes the safest choice of mode. Yeah, that makes sense.

Speaker 2

Yeah. So it can be. Yeah, it could be quite boring if you pick the most probable. One each time.

Speaker 1

Yeah, basically basically. Although like, if you're the accumulation of that is, then it kind of it, yes, this like we sound it sounds different than like what a musician would do because like then when you do it for every note and then that's kind of like a different thing. So I think like the yeah, this is the effect that it can have on the aesthetics. Of the music, let's say.

Speaker 2

Yeah, that's very true, cause that's actually something that I'm looking at in my work is to try and. Because it is a good thing to have repetitive patterns, because it means it's picking up on something and it's learning something. But to try and avoid it, being mundane and and boring kind of patterns. I am. Yeah. So kind of going into that because it's good for long term dependencies. In music, it can span with music, it can span extremely long sequences. Why does this pose a challenge for transformer models specifically? So like you were saying that it could probably if if there's a sequence that's incredibly long and a song that might cause problems is. Do you can you speak to that or? UM.

Speaker 1

Yeah. So I can experience. There are like I haven't seen. Extremely long sequences. It's more that it's missing some like structural. Elements. So I was thinking how to encode these things. Like I I think the the length of the sequences depends on the data quite a lot because like in my I was like. I'll strain it on short sequences, jazz music phrases. And I remember trying to include longer rests as as a character or not and as was was very important. To have less. Generation, but I think the. So the problem with longer sequences like. Is not really that they are long is that they are is that they are, they are not something. Usually when it happens is that they're not something the musician will play is that that they will make any sense. That's the problem. Yeah, so they are long and they might be, they might. Might be some harmony there. Even like they might even like be correct. But the problem is that they don't. They still don't make any sense because like there are no sort of like repetition there. They can they not go up and down, but then they. They don't resemble what the human would play because they are like, for example, they don't know how to like. Groups of nodes are organised for example. Then I am.

Speaker 2

No kiko. When I'm following you.

Speaker 1

Yeah. So for example like. Yeah, like I mean. There are lots of like jazz players, or if you hear like guitar solos for example, all these like very fast, like very long sequences are not those make sense because there are repetition of small patterns. On different points of the so as I.

Speaker 2

Because it's in a scale, I suppose it's. And that kind of thing.

Speaker 1

Yeah, but it's not just the scale. Like the model usually gets the scale, so things are in tune it. They cannot get medium sized patterns and their function, at least my my mother didn't because I didn't say that those were important and it didn't pick up for pick. It up from itself. Let's say.

Speaker 2

OK.

Speaker 1

But I think here once I mean my my project is like three years old now and like Transformers now are like extremely. So probably these things is solved. I haven't checked honestly.

Speaker 2

Yeah, that's the thing I'm finding is I've read papers like a few few months ago, a few weeks ago, but yeah, in the last couple of weeks, there's so many papers have come out already on it. It's, you know, there's a lot of research happening at. The moment. Yeah, that's that kind of answers my question. So that kind of answers the mix as well. What? OK. So what type of data are required for training a transformer model to generate music? There's different types of data that you could use on your transformer, but. You know, is there certain ones that you would aim to go for or there if you were to play pick a data set? Uh, what you'd be looking at for basically?

Speaker 1

Yeah, I guess here is. It depends. It depends on what you're going to use, what music do you want out. Like how you're going.

Speaker 2

To like so for my research. I'm currently working on piano music, so I was going to aim for. MIDI files and for them to be quantised. So that's the way I'm kind of thinking about this books.

Speaker 1

Do you want? Do you want? Sorry. I mean, I don't know. You want to generate? That's the there is. Yeah, there's another maybe thing that can be helpful for you if you want to generate performances. I think that sounds like music like like a human playing a scholar. Do you want to generate that or do you want to generate a score?

Speaker 2

UM, so initially I was thinking of something a bit more human. Like, that's nice to listen to. But after speaking to other people, even though the quantized one would sound more robotic, I think it's easier to. Model on a shorter space of time and there's a bit of more research on it. I think with the human like one uh MIDI files, there's you have to take very small snippets because the timing isn't exactly where you think it should be.

Speaker 1

Yeah. Yeah. Yeah. So yeah, that's probably a good idea, I think. Yeah. The true then, yeah, probably you should go for quantized. I don't know. But polyphonic music that I haven't had because my. My mother was monophonic because I was interested in just. Solos. So I was. I'll give it like a chord, but not another like I even had baseline, but baseline was quantized in fourth, so it like basically one note per. Bit it was not the way. Another moral decline so polyphonic music. I don't know how people do it, honestly. Yeah, but yeah, I would say media contest is good for that. In my case, I was using. I was using. Music, XML and ABC file and media.

Speaker 2

ABC file XML. And maybe so. How different are they those three files and did you amalgamate them? Or how did you go about that?

Speaker 1

Yeah. This. Yeah, that was. The most actually the most like difficult part of the. I think my. I think one data. Set was actually. It was in CSV format. Alright, so yeah. Yeah. So I had to because it was machine transcribed, so I had to quantize it myself in Python.

Speaker 2

Alright, OK, that's that's a lot of work.

Speaker 1

Yeah, it was like. So much work.

Speaker 2

Is there a reason why you went down that route rather than getting like publicly sourced data?

Speaker 1

It is public resources called the database. It's but. And it's really like it's an impressive work. It's like it's a big corpus of transcriptions of just solos. But the problem with that and it's the only public jazz music database.

Speaker 2

All right. OK. Yes.

Speaker 1

That I found. 11 I don't know, probably. Like in the next, in the last three years. There is more, but this one is the the best and the most. The only public one that I. Found yes, a collection of. But if I think like. It Canada must be, I think this the next data set is huge from agenda.

Speaker 2

The sorry, the way to make the data set. Maestro. Yeah, there's a few. There's Lac local. I can't pronounce it. It's something LAK. Something there. There's loads of different data sets and it's there's a lot of. It's a lot of classical piano music is available. There's even, I don't know if you know there's a music app that you can learn how to play the piano, but you can download the MIDI files from there. It's. Muse, something music or something? Well, yeah, yeah, it's. I think that's part of the reason why classical piano is because there's a bit more data available. Yeah. So. I suppose that actually that's kind of goes into the next question is. How did you pre process and represent your music data to be compatible with the model where they're like certain things like outliers and? A certain amount of cleaning to be done for the data to be used in the model and were there things to like kind of look out for because it's music is very different. To like words and you can kind of pick it out a lot easier I suppose.

Speaker 1

Yeah. Yeah, so. Yeah, I think the problem with. Not the problem, but like the. The tricky part of using Transformers is that they are optimised for single sequences and music is polyphonic language language models. So they are built for. Language and languages are not. You speak one letter at a time. There is another. Two at the same time. So you have to find sort of ways to to tell the model that two things are going on at the same time, polyphonic music instead it's like. Different now. So at the same time. So my way was to. Like we build an actor, so for each note I built I I had a square in music, music, XML and for each note I had attacked with different information about the note. So information where each of the note so in meeting so. Can be like cannot. Yeah, like. Whatever like 6456 duration of the Note, and that that I'll try different kind of quantization. So I had for example I had. Yeah, like homemade, like half quarter note, eight note, then I had 1/3 note. I had and then I I was trying to implement like groups groupings of five for example which can be found in jazz. In my desert all this like. Different direction vocabulary. Then I had accord for each for each night I would tell the cord that is like the the cord that displayed underneath. Of the node in the same vector. So it's like each duration 4 nodes that describe the chord. Then I will tell them the next court because India is is also important because this is. Next called is determines where the melody is going to go. And the player who is playing a solo. Knows where all the next court is. Even intuitively it's not even like something that like that is.

Speaker 2

So they're more so. Suppose they know the most common. Chords that go together kind of thing.

Speaker 1

Yeah. Basically, basically. So the next call and then the notes that the play the bass is playing. So what the base? Is playing at the moment. And then the the beat as I said before, so if it's which is either 01/2 or three basically if it's. A4 four piece.

Speaker 2

Is is your music not considered polyphonic polyphonic because you know there's different things happening at one time. I. Know you're saying it's? But because you have the base. And everything, would that not be considered polyphonic?

Speaker 1

Well, no, because I mean I would consider it like that because I'm generating one note at a time. So my mother predicts.

Speaker 2

OK.

Speaker 1

The pitch and the duration one at a time. So given a sequence. What the model does is and the pitch of the next note and the duration of the next note. Then put them together, add them to the sequence, and then predict the next one, and then the next one and then. The next one.

Speaker

OK.

Speaker 1

And the other things are taken from the structure of the piece. So I put them together. And then I before addicting again a source of calculate are we in the next bar or not? If we are in the next bar? Then I change the chords and the base.

Speaker 2

OK. If you're in the next bar, you change the chord in the base. OK. You changed them all together, just depending. On the bar that. You're in it.

Speaker 1

Yeah, so basically the it's like it's like if the model is like following the score you know? So it's following the the score and. It says like. OK, now I am like here in the bar. And like so, the next step is going to be this, and then it generates it updates the data. That the model that it's using to generate. The next node, let's say. I don't think quick polyphonic piano generation. I don't think this is going to be useful.

Speaker

For you.

Speaker 2

OK. Yeah. Sorry, Gwen.

Speaker 1

Yeah, I mean it can be useful. In general, but like I think. Polyphonic music. I would not like how we will do it is I don't think they use any of the of this kind of like. Maybe they have pitch. And duration and that's it. And they do other things, sorry.

Speaker 2

It's just, yeah, it's just, it's just to get more of. An idea? Of the different ways to go about it.

Speaker 1

Yeah, that's really. I think that's a really. Good idea that you have. To if you have to build a model. To interview people, because that's actually like something that I had to. Learn from papers and like. But you cannot read. All the all the papers in the world so. That's actually interesting.

Speaker 2

Yeah. And what I'm finding as well is that I read a lot of papers and I learned a huge amount. And then in a couple of weeks, you'd forget different pieces. And it's only when I interview people. I'm like, yes, that's a very good point. And yeah, they emphasise the pits that I kind of brushed over.

Speaker 1

Yeah, that's that's a lot of. A lot of. Like intuition and knowledge that is like not it's difficult to describe in the paper, but like when you're training the model. You can, at least for me, I would listen to it and say, OK, this is \*\*\*\* is not good enough. So like and that part is difficult to. Describe in a. Paper or like this works. This doesn't. Work like you cannot write everything there.

Speaker 2

Exactly. And there's only so much you can say in in the couple of. Ages, but also I think it's a lot of people have a lot of knowledge on music already, so something that's very secondhand to them might not be to to someone like myself. So like that, are there any challenges or specific considerations when dealing with music data in this context? Yeah. Is there any kind of major challenges that kind of dealing with music data?

Speaker 1

Major challenges. Yeah. I think one challenge is how to represent it because like music is not. If you start from symbolic music, so score. In principle, like even like how you read the score, how the score becomes a sound, so the symbols that are describing the score. Change all the time like they are not. And the the way musicians read the same score. Is very different. Like with piano music, you're kind of that's kind of alright, but just already when you go into jazz for example. The same node can be isolated in a different way, for example. So in just the like. One note is you you read it with the swing, it's called the swing, so the the whole note is red with like. And the piano note would be like that. So that's that's written in the same exact way. So the challenge in general, I think yeah, the challenge in music is how do you? Go from. The encoding. Like musical notation misses a lot of information. That's the. The challenge I think it's it's yeah. It's a description, but it's not. Like it's it's a tool. It's not like itself, so that's that's complicated because then you decide like on which directions if you have a 5. A group of five. Nodes. It's a quilting, so it's like. Is that? Is that so? You have to describe that, but then you have a group of seven notes or like you. Have you can have. Group of 13 notes. And like, that's complicated and different freedoms and and and like like if you look at like like so-called like Western music. Like standard ways of describing it. That's kind of standardised, but. That's not even true for like pop music. Like vocal is the weird things and like guitars. Like like when the bars and. Like and like there are noises. So it's like, yeah.

Speaker 2

It's like No2 musicians play the song the same way, even though there's the same kind of. You need to confront them.

Speaker 1

Yeah. Yeah. Something. Yeah. Yeah, something like that, I would say. Yeah, I guess that's and that's. Challenge. Yeah. I don't know. You can build like convincing representations, but that's maybe. OK. Yeah, I don't. Know, sorry.

Speaker 2

No, that's I think that's a really good point. Yeah, it's definitely something I've read that people have touched on is like you're saying it's more of a guideline of how to play. But it. Yeah, you miss a lot of things unless you know the context. Like you're in jazz. So there would be swing. Notes rather than just. You know you. Need to put my notes.

Speaker 1

The same the same score is interpreted in different ways based on conventions that are not in the score. That's the something like summary. So.

Speaker 2

Based on the genre.

Speaker 1

Yeah, yeah, Chandra and like even like. Like some culture where this is played like. Like you can write the scholar of. A punk song. But that's they don't say anything about what? What is punk and how it's supposed to sound like and how. Yeah, it doesn't have all the information.

Speaker 2

Yeah, that's actually a very good point. And actually I'll go on to a question later on about something similar and. So to keep on to the data set, what does the quality of the generation music improve? Or why does? The quality of the generation music improve with the larger and more diverse training data.

Speaker 1

Yeah, quality. Yeah, I would say it does. I guess it's, it's tricky, it's tricky. It's tricky, I think any. Like there is a point up to which you don't have enough data, so you hear what you generate and it's \*\*\*\*. You don't have enough data that's safe to say that you can like. There is that point where you can recognise that our data is not enough. After you pass this point. It really depends on what you want it to sound like, Mike. And it's it's also this I didn't get. To that point. So I was I was not with my project. I was not satisfied. I I didn't have enough data, I think or my data was very was very, very complicated like like the. Description of free jazz solos where it's like one note for three bars and like the saxophone is screaming, but it's not in the score, for example. And those I kind of try to clean it, clean it and then kind of. Took it away all these things and. Kind of tried to. But still it's very like I would say like. And it doesn't fit in the Convention this data.

Speaker 2

Did you ever. Go down like data augmentation kind of route.

Speaker 1

Yeah, I tried. I tried. The problem is that again like. Like just so that you can do, you can do but and it it kind of helps. If you don't overdo it. It kind of helps with learning the basics, so if you.

Speaker 2

OK.

Speaker 1

If you want your mother to learn what tonality is, for example, to to play in tune. That's all right, because if you augment like how I try to augment is like basically transposition. So I would like I. Would take a melody and transpose it in another key.

Speaker 2

OK.

Speaker 1

And give it to the mother. That's one way. So the model will play more in tune if you because it kind of it has more like evidence that under this court this note is OK basically.

Speaker 2

Yes, OK. Yeah.

Speaker 1

But the the problem is it doesn't claim other things because for example I have like I had saxophone solos and saxophone cannot play in all the keys.

Speaker 2

OK.

Speaker 1

It's then it becomes not realistic, like it doesn't, and then the method is the telephone play, so it's even in the symbols. The melodies that saxophone place are influenced by the fact that saxophone cannot go upper than a certain threshold or lower. Yeah, so they can, they can stay. That's why a saxophone player doesn't play as. A piano player. Because Piano player has more more notes. And that means. Even if they play with only like one out. At a time. That means they play different melodies.

Speaker 2

Yes. Yeah.

Speaker 1

That's so if you want your your data to learn that to learn like a saxophone player. To play like exactly 1. Player you cannot transpose that much, I think. Otherwise, he plans different things. So it's kind of you kind of have to? But that helps with the basics I think. So it's kind of depending what what you think the. Basics are also.

Speaker 2

Yeah, it's how advanced you want it to be. Like your end goal. UM. Yeah, that's interesting. It just shows you how much you need to know about the instrument you're trying to produce the music on because like that, unless you know that you can't just jump up a key like any other instrument on a saxophone.

Speaker 1

Well, I mean The thing is you can't do that if if that's what you're interested in. But it has. So if you want to make music, music means that there are, like, you associate sounds with meaning based on historical conventions. So the way you interpret this music is based on the fact that the saxophone has been played for 200 years and you heard it in this. Song, which you. Liked and blah blah. So if you play a saxophone. Like 1 Octave, Upper Tennessee oxycodone it has. A certain meaning. And a certain feeling. And you're kind of, I guess it's not. Like you have. To be aware of. That it's not really, that's not the point. Like because musicians don't know what they don't care about this theory, all this kind of like, they just place off and think it's cool and it's OK. Yeah, it's kind of different when you do, when you do it on a massive scale and like then you you kind of if you want things to sound a certain way, then it can help to kind of think about all these like associations, let's say.

Speaker 2

Well, I suppose you can really create more better when you have a certain amount of constraints. If because if everything's an option, then you kind of it's hard to know where. To go.

Speaker 1

Yeah, that's a good way to put it. I think that's that's a really good way to. To kind of yeah.

Speaker 2

Yeah, too many options. Is never a good idea, and that's great. And I suppose going into how do you handle outliers in the model during training and could you talk about specific kind of techniques that you used because I know you were talking about that there was. Notes that we're playing for like 3 bars straight and you were taking them out. So how I suppose, how did you go about? That or how? Yeah. How did you handle that?

Speaker 1

Yeah, it's. Sort of like you need to be lucky and. There is, I think my my way was very much kind of like a way of obsessive kind of discovery and like, because The thing is like, I couldn't listen to all of the solos. So I would like, for example, find out that here there is something. That's like this now this been playing for like 3 bars. I discovered that I don't want to remember how and then I decided like I put a rule there like so I selected but now this that are for example like. Yeah, like select many days that are shorter than. I think I had different sizes sequence sizes I call. Them that's I mean scale standard. The transformer models you have sequence size. For example, one thing I would like cut melodies. That are in sequence size of 64 or 32. And longer melodies would be cut, for example, or I was cutting them when I found a long a long rest for example like a like a whole bar rest, even if it's not 64 not I would cut it. Before and part. The rest of the melody, because this is like. Sort of like. So even if it's like a three note melody. A long rest, then, that would be the sequence. If the rest is long.

Speaker 2

Yes. Yeah.

Speaker 1

So that, but that's kind of arbitrary. It was just like my definition of melody.

Speaker 2

Yeah. It's whether you kind of want a long rest or not.

Speaker 1

Yeah. So and then what, which then I thought, what? Which rest is long enough to describe Melody. So is it like a water mattress? Is it like? Have not pressed. Is it a full not rest and that depends on the data set. Yeah, that's like. That depends on the. Data set and. And that's but and in the end it was basically like. Like seeing. The It's also then it becomes also an engineering problem because it was generating with like full no trust cough and not trust. And seeing how it sounds like. And of course, some of them sound like, for example, if I put the the 4/4. Rest I don't know how my. Trust. I think I spent.

Speaker 2

Yes. Yeah.

Speaker 1

My my English music times are not the best.

Speaker 2

It's OK.

Speaker 1

To have a whole mattress then it mean it would mean that there there were not enough melodies that would end with a full trust. So it basically wouldn't capture models that well.

Speaker 2

OK.

Speaker 1

So I needed to to use. The two quarters half not pressed. To describe the matter this. That would generate, I think, different kind of music like bit more like interesting. Yeah, in a way, because there are more separate melodies and it's kind of learning. That's like, OK. This is a sequence like this is what the method is. It's. Yeah, it's basically you're in there this. Is also interesting because this is like. Sort of like. Something that you tell the model that is not even in the architecture, so it could be like a completely unrecognised. Are not described, but it kind of influences a lot how the music sounds like and here you are telling the model what the definition of a melody, what a melody is sort of like in a sideways way. So there are all these like kind of decisions. That doesn't mind what the mothers learns. The mother learns.

Speaker 2

Yeah, I think that's actually a good point to kind of, yeah, it's it's trying to figure out what you want as the melody and kind of. Encourage us that kind of way.

Speaker 1

Yeah, it's like. It's just sometimes it just doesn't work. So that's like a scene between these kind of technical requirements. That's like longer model. It doesn't work, just the just the model. The model is not, doesn't understand anything and what you want it to sound like. So there is like a baseline of like quality and then after that it's like your decision.

Speaker 2

Yeah. Yeah, it's kind of, yeah, getting to that baseline and then kind of tweaking it to improve. UM, so I suppose part of it is we were talking about earlier, uh, trying to trying to avoid getting boring repetition. And uhm, trying to get interesting music is there. And how do you handle like overfishing or under fishing the music? So that you're actually getting. A nice melody or you're getting that the the model works for you. Is there a way of kind of combating that?

Speaker 1

I mean, I've heard some people use this. What's it called? There is a technique. It's called what temperature temperature, yeah. I think it's. Very, very difficult to overfit the transformer model if you have. If you have a big enough data set because. This language is so loud music like they are so complicated that for it to learn everything perfectly. It's kind of like I think. This can only happen if your leader set is very small and then it's like 3. Repeating those 3 songs over and over again.

Speaker 2

What do you consider a small data set like you were saying 3 songs but. You're saying your your data set was small.

Speaker 1

I had 100 something solos in my data set.

Speaker 2

And you didn't run into overfishing in that?

Speaker 1

No, no, no, no. It was like overfitting and that's in that case would be having an output that. The same as the input. So in my case is like I say, improviser this song. So I say to the model generate music or this song and I get one exactly the the same sequence of nodes that I have in one of my of the scores of my data set. That was in the training set. Let's see. That's very difficult to think like you would never like. In my case, I wouldn't. It would be impossible for me to get the same exact solo like that the model. Would have the. Same exact solo. Then the the one that was in the. Training set.

Speaker 2

I suppose flipping that on its head under fishing is that something that needs to be looked out for more so than overfishing?

Speaker 1

Yeah, I guess. I guess. It's a bit strange. Yeah, I guess underfitting is. Yeah, it can be more more of an. Issue when you don't. Have enough variation in the data. And or not. Do you not have enough data? And you don't get enough like. So the model. Doesn't learn what music is, basically. Like it doesn't learn anything.

Speaker 2

Yeah. So.

Speaker 1

I guess that's that's the.

Speaker 2

Selecting your data set probably well is sort. Of all that out.

Speaker 1

Yeah, I think it's.

Speaker 2

So I suppose going into the issue. With transform models is that they're absolutely massive, and they're really computationally expensive to run, and so I suppose an issue that I'm trying to minimise as possible is. UM like trying to optimise your model. Your your transformer model as much as possible with hyperparameters and hyperparameter tuning, but that could take if you're trying to tune your transformer it could take weeks and months to do because it's. So big. So is there a way of like? Making that more efficient that you found. UM. To optimise your model as much as possible.

Speaker 1

I tried I. Mean a little grid search. Yeah, I did a grid search and. Which is basically trying different hyperparameters, different terminations of hyper parameters.

Speaker 2

Did that take a long time to do?

Speaker 1

And training. I mean, I did it. I was training. That's kind of like based was based on intuition in my case because it's like. I have to. Know how many epochs of training is enough? For the grid search. To make sense. So for example, I would say like I could like train over 10 epochs. But then the gate search wouldn't make any sense because the model did not. Converge and it would never converge after 10 epochs, so never. In 100. Would be enough. So I have to know that I guess it takes. It takes quite a lot of like. Heuristics. So kind of like navigating. And it's also quite I mean quite frustrating because you can like train a much bigger model and nothing changes. But then you change some things, some other other parameter and like then it's different. I don't know. I guess my my. Advice is to find. Like start improving and it's also tricky because like before a certain like. Sometimes small models don't converge, like never converge. So I would like update and go to bigger bigger Model 1 by 1 and at some point there is a level where it's not gonna improve anything. Having more more. Yeah, having more weight is not going to improve anything. So that's. Yeah, I guess it's somewhere in the middle. So what about this sort of middle ground other parameter where things sort of work basically?

Speaker 2

Yeah. So kind of going with what's most commonly done and kind of. Hoping for the best that kind of way.

Speaker 1

Basically, yeah, it is frustrating. It's like there is no there. There is no. Mathematic mathematics theory. Yet I think that I know of about how to choose these things. Highly rigorous.

Speaker 2

Yeah, I suppose it's still early days, so there's everyone's still researching and figuring it out. So I realised that we're we're coming up to our time. I just want to briefly. UM. Talk about so when you were evaluating your model on on how well it's done, umm, a lot of people do their research and kind of get surveys to see, uh, what people thought of their music. But I suppose that's kind of you do. That at the very end when. You're kind. Of kind of happy with it yourself. And so it's probably not the most iterative process. So when you are training your data or you're training your model and you were doing the grid search hyper tuning, but is there any other way to kind of uhm? Is there a way of evaluating your model? Without having to do a survey every single time.

Speaker 1

I was using more loss. Like just the loss function. And then I mean in my in. My paper, if you look at. The paper it's like. There are different metrics on different levels of like theories or I was like I was using the language model metrics like Bleu. And the loss. And then I was using this like statistics that are computed. It's called the MG evaluation toolbox. Music generation, relation tool books and they have some code to to test different metrics. Then I also always checking how in tune the notes. Are with the code. And then I had the survey also, but the survey was. Like I mean. It was kind of interesting because, like people were able to kind of recognise and compare what already like it was very, very, very selected, because by the issue when it was always. Like, how do you select? The outputs of the model.

Speaker 2

Yeah, cause when each time we run it, you're going to get a different output.

Speaker 1

Yeah. So basically what people do, what I did is like I selected the best, the, the, the phrases that sounded the best to me. And I think that's the more honest way of doing that. I'm not interested in, like, generating music, but my interest is in artistic applications. So. OK. For me, it's enough that I generate music and that's how I was training so. For me, it's not that the generated music sounds how they want it to sound, so my way of evaluating is like in this case loss was enough to understand how much like if the model, how much the model has converged, and then I would listen to it. And then decide it's OK, this is. Like this is all right or not, all right or. So what my interest is? In this music, I wanted to sound like.

Speaker 2

When you were doing your survey, did you like? How did you find your population? Is it somewhere? That you. Because I'm I suppose I have. I'm wasn't initially thinking of doing it, but the more I kind of look at it and research about it, that seems to be a very common way of evaluating their music. But yeah, so kind of how does he feel kind of? About it. Because it's quite labour intensive for survey.

Speaker 1

So yeah, I think I don't. I'm not sure because I I think. Like honestly, my server was like over like I think like 50 people. I called it online like an app so web application. And people would listen to different phrases and. And vote like trade them from one to five stars. How much they liked it, not even like without saying that this is like from the original data set. This is like from. And then I would tell them like in the end, I will tell them like like which one was generated and which one was from the. Original data set. Yeah, so manipulation. I found them not. Also, they had to say was what's their level of expertise. So if it's like professional musician, I think I had music student and just like. Like amateurs.

Speaker 2

A normal person, yeah.

Speaker 1

Yeah. So I just showed it to the all people I knew. I knew a bunch of musicians, so that was kind of like I got some musicians and then they shared it with their friends and their, like, music school. And like. Yeah. So that was my. It was not very. Rigorous I think. I think this study can be kind of interesting, but if you want to have a rigorous one, it's much more complicated. You need to have so many people and how you build the study. It should be very, very rigorous and how you select melodies should be very, very rigorous. Otherwise it's. I don't think it's very relevant, honestly.

Speaker 2

Umm, because even like you were saying, it's very selective. It's very created to what you kind of want their answers to be. Because yeah.

Speaker 1

Yeah. And also are you the whole thing like how you train it is all like? It's a very, I mean, at least for me it was very like personalised sort of to me. It's like it's not really. I don't know how much is it possible to make something or if it's. Sure, I'm not. Interested in like making like a generalizable system? I think that's not. That's not even if it's possible.

Speaker 2

Yeah, yeah.

Speaker 1

It's not interesting for me. It's not. I don't think it's possible. So this way of evaluating like just like. Yeah, it's not. Maybe it's interesting if it's done properly, but my way is more like let's see. How it goes?

Speaker 2

Yeah. Yeah, I think that's that's a good idea.

Speaker 1

I think so. Sorry. I think one more one much more interesting thing is if you have if you will yourself do an analysis kind of like a musical analysis of the score. So what I what is in my paper is I I kind of took example of musical phrases and compared like so here the money goes like this goes up and down and the core teams are here and like here there is enough paid job. And compared to what the model does. I think that's much, much more interesting because that's so that kind of tells you what did it learn, what did it not learn and like? So a score analysis. More than than survey, but this is my opinion I. Mean I don't. I haven't seen many people do. That because that's arbitrary, it's not science, but it's not considered like.

Speaker 2

Hmm yeah, I suppose, like they're trying to get their primary research and things like that. So they're going for the survey, but that's that's.

Speaker 1

I'm not sure with that, but like, I don't think I don't think it's interesting. It's not very interesting, at least can be best.

Speaker 2

Yeah, because in everyone's paper, all their models work out and everyone loves them. Great. I don't. I don't know. Hold you. Up on your time, because we're gone over, but. I just have. Uh. One or two more questions just to kind of touch on. UM, so we were talking about that a lot of music that is kind of used as Western music in your generation. So how do you address potential biases or limitations and training data to ensure the generation generated music remains unbiased or respectful of various cultural musical traditions? So like you're saying, it's. Music can be very much to that area or the place that it's being played, and is there a way of kind of? Yeah, preventing any biases and limitations to it.

Speaker

This is.

Speaker 1

I don't know, I think. No, I don't. Know how the concept of bias applies to music. Like if you want, that's the thing. Like if you want to build like a generalised system that like considers all the possible music and like then you can talk about bias because like OK, you can say that. Like this music is, or if you claim that you are saying that. But you are like. Like if you're clearing the the music that you generate and this is in. A specific style or something like that. I don't know. I don't know. I think I've seen some some like. Yeah, yeah, I'm not. I'm not sure actually how it applies, because I'm thinking that, like, people also depends on how you're going to use the music, what you do with it like. If you train a data set on folk music. For example and then like. Then you you. Kind of. Yeah, it depends. It really depends on what you do with the time, what you claim this music. Is and like? I don't think that everything is biassed. So we just need to be aware of. The bias and like. Kind of be careful what you do and like be. Respectful of like. Where things come from and like you're not that you're not. Being like. Yeah. Like, like I like I was generating, like, my idea was to generate things that are inspired by jazz music. But I never claimed my model is capable of doing jazz.

Speaker 2

Yes. Yeah.

Speaker 1

And even if I did like. Here is like. Yeah, even if I did like. Honestly like. Jesus is like such a. Big broad term and complicated and then like. No one would think that that the the music comes that comes out of that is just, I don't know, I. Think it's like. That it's different when you come from like. Yeah, I guess I've seen some problematic things like I was doing another project that is not published yet. We were doing an analysis of music generation system that was claiming that that's used for like generating soundtracks. For YouTube videos. Or podcasts. That's basically to not to pay copyright. Yeah, basically, it claims that it's generating hippo, yes. So that's different because probably. The engineers that like. That, that, that decided what is hip hop in the system are probably culturally from a different area than what people hip hop represents, let's say. So yeah, and I guess if you use for example a model that is like, yeah, another ethical thing can be like. All right. Like, let's represent all music with Western notation. Well, that's problematic. Like this is like because and that that's then this is encoded in the models because this is how data gets extracted. So like you analyse music based on the tonality for example. But like tonality is a western concept or like it's not Western but it means different things in different musical traditions. That can lead to like. To like sort of like mocking other cultures, let's say.

Speaker 2

Yeah, well, it's it's not include inclusive I suppose.

Speaker 1

Yeah, but, but I think it's just like. It can lead to like generating. Elevator music in the style of hip hop, which is not like, you know, like this elevator jazz music.

Speaker 2

Yes. Yeah, yeah, yeah.

Speaker 1

And that's what most people think. Not most people will like, but category of people who don't listen to jazz think that jazz sounds like. If you do that in other genres. Yeah, that's. I guess. That's not good, I guess. Like if you think that hip hop is like this, kind of like they're. Like big this. I don't know. Yeah. The risk is to generate here elevator music, I guess, right? Not sure. Like, it depends. I'm not sure what it can lead to, but I guess this is based on how people use it. And like, yeah, it's complicated I guess.

Speaker 2

Yeah, I think that was a good point, like even. It's not even just the data that you're using cause a lot of the data is Western, but also it's the Western notation like it's you can. Music is so diverse and so different depending on the region that you nearly need. Like a model for for each one and. Yeah, that's that's a good point.

Speaker 1

Also, I mean it's also like like like. It's also not considering. A whole like the. Whole kind of. System of practises like but that's like yeah. And until you don't claim that you're doing something that you're not doing, like if you don't claim that this is like that, you're saving the world and like, if you are honest but saying like, OK, this is my experiment, I did the whole thing. And like or I. Did something that it's interesting For these reasons and this is. Yeah, guess you need to. You need to. Sorry. If you are overtime.

Speaker 2

Oh, I'm yeah. Sorry. We're have to keep.

Speaker 1

Like like I'm just talking because this is interesting to me, but like.

Speaker 2

Ohh no I'm. I'm happy to talk if if I don't wanna hold you up. If you've got somewhere to be, that's that's what I'm worried about.

Speaker 1

No, no, no worries I guess. Like I can be here until. 4:30 it's fine.

Speaker 2

I I won't hold you up too long. Yeah, I suppose the very valid points and it's that kind of thing. It's trying to be respectful of that of. UM, yeah, acknowledging that. There's just different types of music in different areas.

Speaker 1

Yeah, I guess, I mean, do you have to be careful? Like I think about cultural appropriation and all this kind of like like if you are like, yeah like? Because it can lead to that. Like to like mocking other cultures and like being like sort of. Yeah. Yeah, I'm doing people, but you have. No idea what hip hop is and. Like how? It what it comes from and like. So you need to be I. Think respectful about that. That's important, but.

Speaker 2

That's the point. Yeah. Yeah. And I suppose it. And it, I suppose it's helpful for yourself that you're actually making the music like, like jazz, like to know the this the swing of this, the music. You know, if if you don't know anything about jazz and you just have dot dot, dot, you know and then to say that that's jazz is. Similar kind of idea.

Speaker 1

Yeah, well.

Speaker 2

Sorry, I'm just trying to figure out. Where I am. So I suppose one of the main things that I'm trying to avoid is mundane, repetitive, repetitive patterns and. I suppose. Are there techniques that you have found that have helped to reduce? Having repetitive patterns that even though it is good to have repetition in music but like to minimise. The kind of boring repetition that you don't always want.

Speaker 1

Yeah, that that was not a problem with my data set because my data set is so crazy that like my just data set was like extremely crazy and complex, and I had the opposite problem that like kind of.

Speaker 2

You were trying to get repetition.

Speaker 1

Yeah, yeah, because my data is. Best way too crazy. I had. I mean I trained. On the full data set and I thought I didn't get that much repetition. I guess. Yeah. No, I just didn't have this problem, so.

Speaker 2

Yeah, I suppose it. Yeah. Like that it it depends on the data set. That you're using. How complex your your data set really is. Is there is just talking about evaluation? I know we kind of touched on the loss function and doing our surveys, but is there like an an evaluation of how effective the self attention mechanism is working itself? Or do you have to evaluate the whole model? Or can you evaluate self attention on its own?

Speaker 1

Yeah, I'm not. I'm not sure that's OK. And you can you can like visualise, they're always visualising the set of attention matrices. But I haven't used them so I'm not not sure.

Speaker 2

That's grand. OK, so kind of my final kind of questions are, why did you decide to focus on specific on the specific field of music generation using transformer models? Uh, what potential do you see this in this field? So kind of, I suppose, in your own research, what kind of probes your interest in? You know, generating jazz music.

Speaker 1

Yeah, in my case. In my case it was kind of like. That's not what my my research is now. So it was like a master project.

Speaker 2

OK.

Speaker 1

I think it was interesting. It was like a personal thing because I was playing guitar and I was like, I wanted a bass. Line like generated baseline. So and I had this like I wanted this kind of like I thought it was. I had this idea of making this like like real time that I could play with the virtual bass player.

Speaker 2

OK.

Speaker 1

So that was my, my my dream. I didn't have time to. To, to. Do that. But I guess I mean. Yeah. And just because I'm like, I studied jazz, I'm like. I'm. I'm not a jazz musician. But I studied jazz. So that's where my most of my like knowledge of theory comes in and like also how it works.

Speaker 2

Hmm. Yeah. So it's just an interest in kind of combining your interest with computer science and. Are there any? Particular resources or references that you found valuable in understanding and advancing your your your research. Is there any kind of? I suppose it's kind of scary with the Internet to find anything, but is there kind of main places that you found very helpful?

Speaker 1

Yeah, yeah, for sure for sure, I mean. I think. There was, I'm sure. I mean, maybe you. Yeah, probably you're aware of Ismere conference?

Speaker 2

I probably won't.

Speaker 1

International check. Check the proceedings of Izmir conference. There's this is for papers. To find like like technical papers.

Speaker 2

Oh yeah, yeah.

Speaker 1

Then the python's documentation is good, I think. And there are some YouTube coding tutorials that I was using.

Speaker 2

Did you find YouTube that helpful?

Speaker 1

Yeah, there is this this channel. You can find it that has. Yes, this guy is good. Can send it to and.

Speaker 2

Will you send me on LinkedIn? Just cause if if it's in the chat I might lose it. Yes, sorry I am somebody else sent me something in the chat and I turned off the chat and then I lost everything.

Speaker 1

This I think he had. Yeah, some like good, good tutorials on the tension mechanism, like and the torch documentation I was basically. Reproducing and encoding models from the touch documentation the Pytorch documentation. And I I was looking on GitHub to saw something that can be interesting is just reproducing someone else's work from the code. So you kind of look at other people's code on GitHub and like. Try to understand how the code works. And like or just like try to. Be like a very. Simplified version of what you're trying to do with, like the most simple version that you can do like just like like one like. See that everything works. And then you update and I think that that was like useful for me.

Speaker 2

There any particular way that you found like finding code the easiest way on GitHub because? Suppose it depends on how people name their their code. Whether you're able to find it or not. But just kind of. Transformer music, something on the lines of that.

Speaker 1

Yeah, I I think I have maybe. Of there was a list of where it is. Yeah, I don't remember exactly. No, sorry I lost. It that's not. Right. No, I guess usually code code that is related to papers is, well, better maintained so. If you find papers that are interesting for you. Than if they have code that's that's better.

Speaker 2

Yeah, rather than people kind of messing around themselves.

Speaker 1

Yeah, yeah. Although I mean it depends. It depends. But that's what I would say. If there is a paper, it's. Yeah, among the all possible options you can find good.

Speaker 2

Papers. Yeah, that's about. That's the best idea. And so this was my last question is would there be anyone else you would recommend to kind of talk to, to have an interview with? That you would that would have an idea of music, general.

Speaker 1

Yeah. Are you interested in symbolic music generation with Transformers, right?

Speaker 2

Yes. Yeah.

Speaker 1

Or not interested, I suppose. Or are you interested in recurrent neural networks and like?

Speaker 2

See, I suppose my plan was to do a transformer and to do the current neural network, but I'm I'm trying to figure out if I have enough time to to do everything so that.

Speaker 1

Because like, OK, there was another paper I used the code from this paper which was called good. It's called bebop net.

Speaker 2

Beatbox now.

Speaker 1

I can send you the repository on GitHub. And this one was quite good. Maybe you can contact the the author of this of this paper. That can be interesting. I haven't. I mean, I haven't thought. I have no idea who she is, but the the paper and the. Code are good. And a person that's very interesting to contact, I think would be props to from Boston, from KTH in Sweden. He is head of the lab that is developing different kind of AI music, but he had a project called Folk Ironman.

Speaker 2

Ohh forgot or and then I think I read that. The other day.

Speaker 1

Yeah. So he's really good. I mean, I know him and he's really, but he's a professor, so he might not have time to talk, but he is. He has a bunch of PhD students who are working on on different kinds of. Different kinds of. Of applications. This is the list of people from his team and sending.

Speaker 2

Oh great.

Speaker 1

Yeah. So the first people are, yeah. Luca also had a project. With that and. Yeah, I don't. Some of them are working on on synthesis models, so not not like audio, not symbolic music, but. Yeah, there are good people to. Talk to and. Hear other people. Let's see.

Speaker 2

Yeah, I suppose initially when I was thinking of doing interviews because we've been encouraged to do interviews, I was kind of going out into anyone that had music and transformer in their LinkedIn description. But I think like you're saying it's probably easier to target. Am if there's a peak where you like to kind of target their code and interviews that way.

Speaker 1

Yeah, yeah, yeah. I think there is a big club at. In London. What's the name? I don't remember which university? Anyways, yeah, I think it's also. People respond. Barcelona are working on that. A lot of people from the music group there.

Speaker 2

OK.

Speaker 1

Yeah, I know that people that are working on on this kind of Transformers, I think so. Yeah, maybe you can Google them.

Speaker 2

Yeah, that's correct.

Speaker 1

And rising music technology. Yeah, I think. Yeah, I think that's that's some people. That you could contact.

Speaker 2

That's brilliant. Thanks a million for that. I think we've we've gone through all my questions and I'm sorry for running overtime. But I just want to say thank you so much for your time. And like I learned so much in this hour and a half than I probably would if I've been searching for it on the Internet on my own. So.

Speaker 1

Yeah, that's what I think.

Speaker 2

Yeah, definitely. So thank you so much. And yeah, for replying to my message. I've been emailing. A lot of people and getting replies, it's it's like Christmas every time. So yeah, I just want to say thank you very much. And yeah, enjoy your your evening stuff. Annoying, you know.

Speaker 1

Yeah, that's good. Well, it's a. Start with your project.

Speaker 2

Thank you very much. Let's see, by the way.

# Glossary

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Lead sheets | Simple musical sheets containing only the main parts of a song like the melody, lyrics, and chord symbols. |
| MIDI data | Information sent between electronic musical devices that includes details like which notes to play, how long to play them, how loudly, and other control signals. |
| Music elements | The basic building blocks of music, including things like how high or low a note is (pitch), the beat or rhythm, how notes sound together (harmony), the main tune (melody), how the music feels (texture and dynamics), and the overall structure (form). |
| Multi-track | A recording or song made up of several separate parts or layers that can be adjusted independently. |
| Nucleus sampling | A way of picking words or notes in computer-generated music or text where choices are made from a smaller, more focused selection rather than the whole range of possibilities. |
| Pitch | How high or low a sound is, determining if it sounds high-pitched like a bird chirp or low-pitched like a deep voice. |
| Polyphonic music | Music that has multiple different melodies playing at the same time, like a harmony in a choir or a melody with accompaniment. |
| Symbolic music | Music represented using symbols like musical notation, rather than as actual sounds. |
| Temperature parameter | A setting used in creating music or text with artificial intelligence, controlling how random or unpredictable the output is. |
| Tempo | The speed at which music is played, measured in beats per minute, determining if a song feels fast like a race or slow like a stroll. |
| Time Signature | A notation in sheet music indicating how many beats are in each measure and which type of note gets one beat, helping musicians keep track of the rhythm. |
| Top-k | A method used in computer-generated music or text where only the most likely options are considered at each step, rather than all possible choices. |
| Overdriven guitar | A technique used on electric guitars to make them sound distorted and crunchy by increasing the volume or gain beyond normal levels.  Top of Form  Bottom of Form |

Table 2: Glossary of Terms