

**Software Engineering Department  
ORT Braude College**

**24-2-D-20**

**GENERATING MUSIC WITH SENTIMENT USING**

**TRANSFORMER-GANS**

**Ariel Dahan 315682054**

**Niv Malka 314730995**

**Supervisors:**

**Dr. Renata Avros**

**Prof. Zeev (Vladimir) Volkovich**

**Table of Contents:**

1. **Introduction** ……………………………………………..…………… 4
   1. Motivation ………………………………………………………… 4
   2. Project Purpose …………………………………………………… 5
   3. Current Solutions ………………………………………………….. 6
      1. MusicLM ………………………………………………….. 6
      2. MuseNet …………………………………………………… 6
      3. Additional Solutions ………………………………………. 7
      4. Comparison of Existing Solutions to the Selected Model … 7
2. **Background** ………………………………………………...…………. 8
   1. Evolution of Music Generation Technology ………………………. 8
   2. Emotion Representation and Sentiment Analysis …………………. 8
   3. The Role of Deep Learning and Neural Networks ………………… 9
   4. Combining GANs and Transformers ………………………………. 9
3. **Challenges in Engineering Development** ……………………………. 9
   1. Complexity of Emotional Representation ………………………….. 9
   2. Cultural Differences in Emotional Perception of Music ………...… 10
      1. Data Diversity …………………………………………….... 10
      2. Multidimensional Emotion Representation ……………...… 10
      3. Transformer-GAN Architecture ………………………….... 10
4. **Project Implementation Overview** …………………………………... 11
5. **Model Architecture** ………………………….……………………….. 11
   1. Transformer-GAN Architecture …………………………………… 11
      1. Transformer ……………………………………………….. 11
      2. GANs (Generative Adversarial Networks) ………………... 12
   2. Generator and Discriminator Models ……………………………... 12
      1. Generator ………………………………..…………………. 12
      2. Discriminator ……………………………………….…….... 13
      3. Interaction Between Generator and Discriminator ………… 13
   3. Training Process …………………………………………………… 14
      1. Loss Function ……………………………………………… 14
      2. Iterative Training …………………………………………... 15
      3. Other Techniques ………………………………………….. 15
   4. Integration of Emotion Representation ……………………………. 16
      1. Valence and Arousal Metrics ……………………………... 16
      2. Multi-Dimensional Emotion Representation ……………… 16
      3. Sentiment-Conditioned Training ………………………….. 16
   5. Model Architecture Diagram ……………………………………… 16
   6. Challenges and Solutions ………………………………………….. 17

5.5.1 Algorithmic Challenges ……………………………………... 17

* 1. Model Optimization ……………………………………………….. 18
     1. Balancing Generator and Discriminator …………...……… 18
     2. Capturing Long-Range Dependencies …………………….. 18
     3. Techniques to Address Training Challenges ……………… 19

1. **Data** ………………………………………...…………………………. 19
   1. Data Utilized ………………………………………………………. 19
      1. AILABS17k Dataset ………………………………………. 19
      2. EMOPIA Dataset ………………………………………….. 20
   2. Data Preparation …………………………………………………... 20
   3. Data Limitations ……………………………………….………… 20

6.3.1 Computational constraints …………………………………. 21

1. **Evaluation and Validation** …………………….…………………… 21
   1. Metrics …………………………………………………………… 21
      1. Emotion Accuracy ……………………………………….. 21
      2. Musical Coherence ………………………………………. 21
      3. Subjective Evaluations …………………………………... 21
   2. Validation Process ………………………………………...…….. 22
      1. Comparative Analysis with Human-Composed Music …. 22
      2. Listener Feedback on Emotional Impact ………………... 22
      3. Iterative Refinement …………………………………….. 22
2. **Testing Plan** …………………………………………………..……. 22
   1. Optional: Integration with User Interface (UI) …………………. 24
3. Visualizing the Music Generation Process …………………………. 24
   1. Sequence Diagram ……………………………………………… 24
   2. Class Diagram ………………….……………………………….. 24
4. **System Workflow** …………….……………………………...…….. 25
   1. Detailed Process Flow ……………………………………… 25
      1. Data input ………………………………………………. 25
      2. Data preprocessing ……………………………………... 25
      3. Model processing ………………………………………. 25
      4. Post-processing ………………………………………… 26
      5. Output generation ……………………………………… 26
5. **Software Architecture Overview** …….………………………….. 26
   1. Input Layer ………………………………………………… 26
   2. Processing Layer …………………….…………………….. 26
   3. Output Layer ………………………….…………………… 26
6. **Next Steps** …………………………………...……………………. 27
   1. Development Plan for the Coming Semester ……………… 28
   2. Testing and Iteration Plan for the Upcoming Semester ….... 29
7. **References** …………………………………………………..……. 30

**Abstract**

Recent advancements in **deep learning models** have significantly impacted music generation; however, generating music that authentically mirrors human emotions remains a complex challenge. This project proposes a **Model based on Transformer-GAN** architecture, aimed at generating music that reflects specific emotional states, such as **valence** (positivity/negativity) and **arousal** (intensity). By leveraging the **sequence modeling** capabilities of Transformers and the **realism-enhancing** strengths of Generative Adversarial Networks (GANs), our approach seeks to produce music that embodies targeted emotions and adapts dynamically to shifts in emotional input.

Our dual-stage model is designed to **capture long-term dependencies** in musical sequences while ensuring high-quality, authentic outputs. This precise emotional conditioning and adaptive music generation address key limitations in existing models for music generation.

With potential applications in areas like entertainment, therapeutic interventions, interactive media, and personalized content, this model enhances user engagement by generating emotionally resonant music. Moreover, this project contributes to the growing field of emotion-based computational models by providing new methodologies for music generation. Preliminary evaluations indicate the promising ability of the model to align generated music with human emotional perceptions, offering valuable insights for the integration of **Transformer-GAN models** in creative domains.

1. **Introduction**

The development of a music generation system, which is guided by emotions, marks an exciting advancement in deep learning and creative technology. Music, as we know, has a unique power to convey a range of complex emotions, acting almost as a universal language that people everywhere connect to. Aligning with a music generation with specific feelings opens up a lot of new opportunities—not only for various industries, but also for personal uses. This project aims to push the boundaries of how we experience and interact with music in an emotionally resonant way.

*1.1 Motivation*

Music has its own way to present the kind of feeling one is having either happy, sad, excited, or remindful feelings are essential components of experiences. Just consider an environment where this system can actually create tracks of one mood, for another – this can actually be of great use in therapy sessions, showcasing how people can build soundscapes that are dependent on the kind of feeling that one might have need of and might help heal them emotionally or even alleviate stress. This type of technology can also give rise to service that automatically adapts playlists based on current mood of the user, thus playing tunes that would fit the exactly felt mood. Music generation for emotions is not limited to self-only, but entertainment has large opportunities in the future. Thus, the creators attract the audience even more if using the soundtracks that are closer to the feelings in movies, games or ads. Furthermore, such technology may enhance human computer relations, such as through music prescriptions adapted to the mood of an individual or in interactive or virtual scenarios when the track list changes in response to the plot mood. The integrations involved could revolutionize how people interact with content, and in a way, bring more value to the information they consume and share online. This project is also being extended to various other creative and acoustic prospects. This was a new way that composers and musicians could engage with Deep Learning Systems and asked what it even means to ‘translate’ emotions into music. It is a goal of this initiative to bring information as to how music affects emotions and how in turn, help in the composition of music as well as in the scientific field of analyzing music.

*1.2 Project Purpose*

Our project employs a Transformer-GAN model to compose music with desirable qualities of valance and arousal, that is, positivity/negativity and intensity, respectively. The objectives include: generating music that can express emotions like joy, sorrow, enthusiasm, or serenity through sequence modeling of the transformer framework and realistic music generation by GAN. The proposed system will generate smooth transitions between different emotions and, at the same time, will enable creating emotion-based user profiles to personalize real-time musical accompaniment according to situation and individual preferences. Evaluation aids, and testing techniques are also being pioneered to ascertain correctness and calibre of the emotional content of the music. Furthermore, we are desiring to bring the solution to entertainment, therapy, and interactive media industries for improving user- emotional interaction with musical experiences.

*1.3 Current Solutions*

In recent years, several advanced models have made significant progress in the field of deep learning-driven music generation, incorporating emotional elements and pushing creative boundaries. Despite these achievements, current solutions still have some limitations that our model seeks to address.

*1.3.1 MusicLM*

MusicLM is a recent music generation that Google Research has developed to generate music based on descriptions. By following precise rules concerning mood and style of the piece, it produces the music using highly sophisticated mathematical equations. Main advantages: text and image input processing in order to generate music, determining music continuation based on the input style with references to the high quality of output while warning against its usage for professional purposes including music scoring for films and creating specialized playlists. Nevertheless, the application of MusicLM is most useful in those cases where it is important to achieve a given emotion, for example, in creating a library of movie scores or a library of highly personalized music works. Thus, it is prone to inability to model and follow the dynamics of and the course of emotion, as well as possible adaptation there of what may be an issue in interactive or real-time as opposed to user-initiated settings.

תמונה שמכילה צילום מסך, טקסט, עיצוב

התיאור נוצר באופן אוטומטי

Figure 1: MusicLM interface visualization

*1.3.2 MuseNet*

MuseNet is another model developed by OpenAI that has massive capabilities for creating music that spans different instruments and genres. The program does not directly incorporate emotions, however, its capacity to create linguistically rich and stylistically diverse texts helps it suggest emotions indirectly. They include polyphonic composition which allows more than one melody and instruments and harmony, style versatility which can create all style of music be it classical or modern like pop and jazz, long term structure needed to create extended musical works which is important when creating emotional flow. It is worth using MuseNet for generating music in which the affective content is signified by style and orchestration, as it is not directly manipulated. Nonetheless, the flexibility of MuseNet in achieving highly nuanced musical arrangements, it is not very clear how to regulate for affective valence. Its use of style and genre to indicate emotions may not always coincide with the needed ‘emotional content’ for an application, occasionally producing rather ornamental music which may nonetheless fail to convey the desired affective density.

תמונה שמכילה טקסט, צילום מסך, שעון, עיצוב

התיאור נוצר באופן אוטומטי

Figure 2: MusNet interface visualization

*1.3.3 Additional solutions*  
Other applications, like AIVA and Jukedeck, have also contributed to music generation by incorporating some level of emotional expression. However, like MusicLM and MuseNet, they struggle to offer precise control over the emotional tone of the music. This challenge is especially noticeable when music needs to shift in response to changing emotions, where accurate emotional alignment is key.

*1.3.4 Comparison of existing solution to the selected model*

Our proposed Transformer-GAN model introduces a promising approach to emotion-conditioned music generation, focusing specifically on integrating emotional metrics like valence (positivity/negativity) and arousal (intensity) directly into the music generation process. This allows the model to reflect a range of emotional states more closely, addressing some limitations of models like MusicLM that rely primarily on user-provided text prompts. One area where our model aims to make a difference is in handling dynamic emotional transitions, which can be challenging for existing models such as MusicLM and MuseNet. By combining the sequence modeling ability of Transformers with the realism feedback from GANs, our model seeks to generate music that evolves in a way that feels more naturally aligned with changing emotional inputs. In contrast to approaches that rely on general prompts or stylistic cues, we use quantifiable metrics (valence and arousal) to give more direct emotional conditioning. This provides a reproducible basis for creating music that aims to capture specific emotional shifts, allowing users to guide the emotional flow of the compositions more systematically. Additionally, our model explores real-time adaptability to emotional changes, which could enable applications where music responds dynamically, such as in adaptive soundtracks for gaming or personalized playlists. While still in the development and testing phase, this model is a step toward more interactive music experiences that adapt to users' evolving emotional contexts. In summary, although models like MusicLM and MuseNet have laid valuable groundwork in deep learning-based music generation music generation, our Transformer-GAN model offers an exploratory step toward greater emotional flexibility, aiming to provide music that reflects a more nuanced and adaptive emotional range. This project serves as an academic investigation into these possibilities, contributing to ongoing research in the intersection of deep learning and creative technologies.

1. **Background**

Understanding how emotion-driven music generation works starts with a look at the main technologies and methods that make it possible.

*2.1 Evolution of music generation technology*

Music generation technology has come a long way, evolving from simpler, rule-based systems to advanced deep learning models. Earlier methods, like rule-based algorithms or Markov chains, could create predictable and organized music patterns, but they often lacked the creative touch that makes music feel “human.” With deep learning, models can now capture more complex musical patterns, bringing us closer to generating music that reflects genuine creativity and variety.

*2.2 Emotion representation and sentiment analysis*

The ability of music to convey emotion is central to its impact on listeners. In AI models, emotions are often represented along two main dimensions: valence (whether the emotion is positive or negative) and arousal (how intense the emotion is). By focusing on these dimensions, sentiment analysis helps models generate music that aligns with specific emotional states, making it particularly useful in interactive applications where the music may need to adjust in real-time to reflect changing moods or scenes.

*2.3 The role of Deep Learning and Neural Networks*

Deep learning has been essential in music generation, enabling models to pick up on subtle patterns in musical data. Neural networks—especially RNNs and CNNs—are highly effective for analyzing and reproducing sequences, making them well-suited for music, which has its own natural flow and structure. These models allow us to move beyond simple repetition, letting AI create music that feels rich and dynamic.

*2.4 Combining GANs and Transformers*

Combining Generative Adversarial Networks (GANs) with Transformers has proven to be highly effective for music generation. GANs use a feedback loop, with one part generating music and another part evaluating its quality, which helps produce more polished and realistic outputs. Transformers, on the other hand, capture the relationships between musical notes and patterns over time, which is important for creating music that sounds coherent and flows naturally. Together, they create a system that can produce music that is both emotionally expressive and structurally sound.

### **Challenges in engineering development**

Creating a system for emotion-conditioned music generation presents some unique challenges. Here are a few of the main issues:

*3.1 Complexity of emotional representation*

Emotions are inherently complex and often subjective. They can’t always be easily categorized because they span multiple dimensions, like **valence** (whether an emotion feels positive or negative) and **arousal** (the intensity of the emotion). Additionally, people experience and interpret the same emotions in different ways. Designing a model that can accurately capture this complexity and variability in emotional expression is no small task—it requires finding a balance between general emotional patterns and individual variations.

*3.2 Cultural Differences in Emotional Perception of Music*

Emotions conveyed through music often differ widely between cultures. Musical elements like melody, harmony, and rhythm can have unique emotional meanings depending on the cultural background of the listener. This cultural variability makes it challenging to develop a model that generates music with emotional depth that resonates universally. Instead, the model has to learn to adapt its outputs to appeal to diverse cultural contexts. To manage the emotional complexity and cultural variability in music generation, we adopted several strategies:

***3.2.1 Data diversity***

We trained the model on a variety of datasets, like AILABS17k and EMOPIA, which include a broad spectrum of musical styles and emotional tags. This diversity helps the model generalize, making it more adaptable to different cultural contexts and emotional expressions.

***3.2.2 Multidimensional emotion representation***

By including metrics for **valence** and **arousal**, the model gains a more comprehensive view of emotional expression. This approach enables the system to capture a wider range of emotions, helping it generate music that reflects subtle shifts in feeling more accurately.

***3.2.3 Transformer-GAN architecture***

The use of Transformers allows the model to capture long-term dependencies in music sequences, while GANs enhance the realism of the generated outputs. Together, these components enable the model to produce music that is both emotionally resonant and adaptable to different cultural interpretations, striking a balance between emotional complexity and cultural sensitivity.

1. **P****roject implementation overview**

The Project Implementation Overview presents the entire workflow, from data collection to delivering the final music output to the user.

Optional

Figure 3: Project implementation overview

1. **Model Architecture**

*5.1 Transformer-GAN Architecture*

*5.1.1 Transformer*

The Transformer model plays a central role in generating music by handling sequence modeling, which is especially important in capturing the long-term patterns and relationships found in music. Using self-attention, the Transformer analyzes how different parts of a musical sequence relate to each other, allowing it to grasp intricate musical structures. This setup helps the model create compositions that develop logically and naturally, keeping elements like harmony, melody, and rhythm in sync.

*5.1.2 GANs (Generative Adversarial Networks)*  
To make the generated music feel more realistic and emotionally aligned, we integrate GANs into the architecture. This is how it operates: the Generator creates music sequences guided by the structure the Transformer provides. These sequences are then evaluated by the Discriminator, which acts as a quality check, determining whether the music sounds authentic and matches the intended emotional tone (using metrics like valence and arousal). Through this back-and-forth process, the Generator learns to refine its outputs, gradually producing music that becomes harder and harder to distinguish from real, human-composed pieces, thanks to the constant feedback from the Discriminator.

A diagram of a block diagram

Description automatically generatedA diagram of a process

Description automatically generated*5.2 Generator and Discriminator Models*

Figure 4: Generator architecture diagram

Figure 5: Discriminator architecture diagram

*5.2.1 Generator(a):*  
The Generator is responsible for creating new musical sequences based on specified conditions like emotional states (e.g., happiness or sadness). It works as follows:

Embedding layer: Converts the input sequence (notes or musical tokens) into high-dimensional vectors. A Positional Embedding is also added to retain the order of elements, essential for maintaining musical structure.

Attention block: Here, the Transformer architecture uses self-attention to capture relationships between sequence elements, even those far apart. This capability allows the Generator to maintain coherence and recognize recurring themes or motifs across the composition.

Fully Connected (FC) layer: Refines the processed information from the attention mechanism to enhance musical detail.

Conditioning and output: Integrates Condition Embedding, enabling the model to tailor its output to specific emotional inputs. The final musical sequence reflects both the original context and the chosen emotional state.

*5.2.2 Discriminator(b):*

The Discriminator evaluates whether the music generated by the model is realistic or authentic. It operates as follows:

Input embedding: Similar to the Generator, it embeds the sequence (real or generated) in a high-dimensional space.

Positional and CLS Embedding: Positional Embedding helps with sequence order, while CLS (classification) Embedding helps differentiate input categories.

Attention block: Examines relationships within the musical sequence to assess coherence and consistency, identifying long-range dependencies that suggest realism.

Fully Connected layer and prediction maps: A Fully Connected layer refines the sequence analysis, generating two types of predictions: Local and Global Prediction Maps. These provide judgments on the smaller, detailed sections of the music and the overall structure, respectively, enabling a comprehensive assessment of authenticity.

*5.2.3 Interaction between Generator and Discriminator*  
The integration of Transformer and GAN architecture brings a unique advantage to music generation:

Sequence modeling: The Generator leverages the self-attention of the Transformer to grasp complex, long-range dependencies in music, producing sequences that are coherent and dynamically varied.

Adversarial process: The Generator and Discriminator engage in an adversarial setup, where the Generator creates music aiming to sound realistic, while the Discriminator assesses its authenticity. Through this iterative process, the Generator refines its output to better "fool" the Discriminator, achieving increasingly realistic results.

Emotional conditioning: By conditioning the output on emotional inputs (such as valence and arousal), the Generator can produce music that not only sounds authentic but aligns with specific emotional contexts, a valuable feature for adaptive or personalized music applications.

Realism and feedback loop: With each iteration, the Discriminator provides feedback to the Generator, helping it improve both the quality and emotional accuracy of its musical sequences. This feedback loop enhances the ability of the Generator to produce music that closely resembles human compositions.

The Transformer-GAN architecture combines the strengths of Transformers for capturing complex musical sequences with the iterative refinement of GANs, producing music that is both emotionally expressive and structurally coherent. The Transformer models long-term dependencies, while the GAN enhances the realism, making this architecture highly effective for generating quality, emotion-driven music.

*5.3 Training process:*

The training process of the Transformer-GAN model is structured to align generated music with realistic sound and intended emotional cues. This involves fine-tuning both the Generator and Discriminator, using specific loss functions and a structured training workflow.

*5.3.1 Loss function*

The Generator and Discriminator networks are trained with respect to each other, meaning that after Discriminator evaluates the quality of fake samples generated by Generator, the Generator improves the quality of fake data depending on the result of the Discriminator. The Discriminator also uses the binary cross-entropy loss function to determine whether a sequence is real or generated and optimizes it step by step to better distinguish between the two. On the other hand, in the reverse direction, the Generator aims at reducing the capacity of the Discriminator to reject its results as fakes, thereby boosting the potential of the Generator to create better music. For the purpose of ensuring that the model aligns to the correct emotion, an added loss function comes in to quantify the distance between the intended emotion (as described by both the valence and arousal) and the resultant music. This is often done via Mean Squared Error (MSE) which helps in matching the emotional tone of the generated outcome with the required emotional input thus enhancing the emotional correctness of the music.

*5.3.2 Iterative training*

The Generator and Discriminator are trained sequentially to produce musically correct and emotionally engaging music. Within the training loop, the Generator generates sequences with emotional input, and the Discriminator decides whether the sequence is real or fake. Since Discriminator delivers a high loss to the Generator when it successfully identifies fake data, the Generator learns to modify its outputs to resemble realistic and emotionally appropriate data. On the other hand, if the Discriminator believes that the generated sequences are real, it gets feedback to correct the ability of the Discriminator to distinguish between real and fake music. Furthermore, during the training of the Generator, it possesses the essence of not only fooling the Discriminator but also to generate images that correspond to prescribed emotional constraints. This emotion alignment loss is then propagated backwards in order to get generation of music that embodies the intended mood.

*5.3.3 Other techniques*

Optimization technique: The training process employs the Adam optimizer for both networks, allowing adaptive learning rates for efficient convergence. To avoid issues like mode collapse, additional strategies such as learning rate scheduling and gradient clipping help maintain stability and balance between the Generator and Discriminator.

**Gumbel-SoftMax technique:** The Gumbel-SoftMax method addresses the need for discrete outputs, such as musical notes, which traditional SoftMax doesn’t handle efficiently in sequence generation. By providing a differentiable approximation of discrete sampling, Gumbel-SoftMax ensures smooth gradient flow, essential for producing coherent and structurally stable music sequences. Unlike standard SoftMax or reinforcement learning approaches, Gumbel-SoftMax is better suited for tasks involving discrete choices. It bypasses the complexity of reward-based reinforcement signals, making it easier to integrate within the Transformer-GAN framework while supporting stable training and effective sequence generation.

*5.4 Integration of emotion representation:*

*5.4.1 Valence and arousal metrics*

These metrics play a crucial role in conditioning music generation, guiding the model to align with specific emotional states like joy, sadness, or excitement. By leveraging both valence and arousal, the generated music reflects the desired emotional tone, making each piece contextually and emotionally resonant.

*5.4.2 Multi-Dimensional emotion representation*

Employing a multi-dimensional approach to emotion enables the model to capture complex emotional nuances, providing more refined control over the emotional content of the generated music. This structure allows the model to respond to both valence and arousal inputs, ensuring an output that resonates with the emotional intent of the user.

5.4.3 Sentiment-Conditioned training

During training, valence and arousal metrics guide the Generator in producing emotionally aligned music. The Discriminator uses these metrics to evaluate and provide feedback, ensuring that each generated sequence aligns with the intended emotional outcome.

*5.5 Model Architecture Diagram*

This diagram provides a structured view of the workflow and components within the emotion-conditioned music generation system. The architecture is based on a Transformer-GAN framework, combining the strengths of both models to achieve musically coherent and emotionally resonant outputs.

A diagram of a process

Description automatically generated

Figure 6: Model architecture diagram

*5.6 Challenges & Solution:*

The development of the Transformer-GAN model for emotion-conditioned music generation involved overcoming significant challenges related to model optimization and ensuring coherence in musical sequences.

*5.6.1 Algorithmic Challenges*

During GAN training, exposure bias was a concern, as the predictions of the model became heavily influenced by its own generated outputs rather than real data. This created a feedback loop that could reduce the quality and diversity of the music. To address this, the Gumbel-SoftMax technique was implemented, providing a more precise sampling method for handling discrete data, such as musical notes. Additionally, scheduled sampling was applied to gradually incorporate real data during training, reducing the bias and enhancing the ability of the model to produce realistic and emotionally aligned sequences. Moreover, balancing the Generator and Discriminator during adversarial training was another challenge, as imbalances could lead to issues like mode collapse (where outputs become repetitive) or vanishing gradients, which stifle learning. To counter these issues, techniques such as gradient penalty, adaptive learning rates, and early stopping were used. These methods ensured stable training, preventing the model from getting stuck in repetitive patterns or failing to generate varied, high-quality music.

*5.7 Model Optimization*

In training the Transformer-GAN for emotion-conditioned music generation, maintaining an effective balance between the Generator and Discriminator and capturing long-range dependencies in music presented challenges. Here is how these were addressed:

*5.7.1 Balancing Generator and Discriminator*

A core issue with GANs is achieving a balanced improvement rate for the Generator and Discriminator. If one becomes too powerful, it can hinder the progress of the other, leading to ineffective training. For instance, if the Discriminator quickly detects all generated sequences as fake, it restricts the learning ability of the Generator. Conversely, if the Generator becomes too strong, the Discriminator struggles, which can lead to “mode collapse,” where the Generator repetitively produces similar outputs. To address this, learning rate adjustments were made for both networks to achieve balanced training progress. Batch normalization was used to stabilize learning, and label smoothing (applying values like 0.9 instead of 1 for real samples) was introduced to reduce Discriminator overconfidence, providing more robust adversarial training.

*5.7.2 Capturing Long-Range Dependencies*

The self-attention mechanism of the Transformer is excellent for recognizing long-range dependencies. However, balancing short-term transitions and the overall structure of the music was challenging. Focusing too much on immediate patterns can cause loss of musical flow, while focusing too heavily on long-term structure may impact the smoothness of note transitions. Positional encodings were refined to help the Transformer capture both local and global dependencies in the music. Additionally, gradient clipping was applied to keep gradients within an appropriate range, which helped stabilize training and maintain musical coherence.

*5.7.3 Techniques to address training challenges:*

Learning rate tuning: By adjusting and scheduling the learning rates, both networks progressed without oscillations or instability.

Gradient clipping: This technique was applied to prevent gradient values from becoming excessive, which supports the stability of the model, especially in lengthy music sequences.

Label smoothing: This method helped avoid overconfident predictions from the Discriminator, making the adversarial training process smoother and more stable.

1. **Data**

For this project, selecting and preparing a robust dataset is essential to train a model capable of generating emotionally resonant music. After careful consideration, we chose theAILABS17k and EMOPIA datasets due to their diversity and depth in musical and emotional content.

*6.1 Data utilized*

*6.1.1 AILABS17k dataset*

AILABS17k dataset includes over 17,000 hours of automatically transcribed piano covers of pop songs in MIDI format. It offers a broad range of musical patterns and emotional content that allows the model to learn varied musical structures.

*6.1.2 EMOPIA Dataset*

EMOPIA dataset Designed specifically for emotion-based music generation, the EMOPIA dataset contains musical passages labeled with valence and arousal metrics. These emotional tags enable the model to understand the correlation between musical elements and emotional states, allowing for the generation of music that aligns with specific emotional inputs.

*6.2 Data Preparation*

All MIDI files are cleaned and normalized to ensure consistency. The emotional labels are aligned, and the data is converted into the proper format for model input. Proper preprocessing is critical to ensure accurate emotional representation during training. Data augmentation techniques such as pitch shifting, time stretching, and noise addition are employed to increase the diversity of training data. This allows the model to generalize better, producing varied musical sequences that capture a range of emotional tones.

In the upcoming semester, we aim to leverage these datasets to further train the model, enhancing its ability to generate music that aligns with specific emotional states. With these rich and varied data sources, we seek to refine the capacity of the model to learn from both musical patterns and emotional metrics, enabling more accurate and emotionally resonant outputs.

*6.3 Data Limitations*

Realizing that there are limitations in terms of the training data, methods such as pitch shifting, time stretching, and adding noise to the audio samples were used. These techniques modified the dataset, making it easier for the model to learn and do well on unseen data. Transfer learning was applied where necessary to increase performance with minimal data. From the prior knowledge, we were able to improve on it therefore limiting the amount of training that would take place from the scratch thus cutting short the time as well as computation time.

*6.3.1 Computational constraints*

Due to high computational complexity of deep learning models the cloud resources like Google-Colab and AWS were taken into consideration. There was sufficient GPU availability and the level of control required for many large scale training tasks on these platforms. Also, it is worth using the strategy for optimum coding so as to use the resources efficiently. Hence, it is possible to have successful loading of computational tasks using optimal libraries, parallel processing as well as the avoidance of extra calculations within training processes.

1. **Evaluation and validation**

The effectiveness of the model in generating emotionally accurate and musically coherent sequences was assessed through both objective metrics and subjective evaluations, following the completion of its training. Here is an outline of the evaluation and validation process:

***7.1 Metrics:***

***7.1.1 Emotion Accuracy***

The primary goal is to align the emotional content of the music with specific inputs, such as valence (positivity/negativity) and arousal (intensity). Emotion alignment metrics measured how closely the generated music reflected the intended emotional states, with Mean Squared Error (MSE) used to compare target and actual valence/arousal values in the output.

***7.1.2 Musical Coherence***

The ability of the model to produce musically consistent sequences was measured using Cross-Entropy Loss, which assesses how well the generated music aligns with the statistical distribution of the training data. Additionally, Perplexity scores evaluated how accurately the model could predict subsequent notes or chords, helping to ensure coherence both within and across sequences.

***7.1.3 Subjective evaluations***

To complement quantitative data, feedback from human listeners was gathered on emotional accuracy and overall quality. Listeners rated the emotional tone, coherence, and realism of the music on a scale, allowing us to compare their feedback with the intended outputs of the model.

***7.2 Validation process:***

***7.2.1 Comparative analysis with Human-Composed music***

Generated music was compared with human compositions across various genres to assess realism and stylistic alignment, ensuring that the model produced authentic and genre-appropriate music.

***7.2.2 Listener feedback on emotional impact***

Human participants provided insights on whether the generated music matched the intended emotions. For instance, if listeners often rated happy compositions as neutral, adjustments were made to improve the ability of the model to accurately convey emotions.

***7.2.3 Iterative refinement***

Both metric-based results and subjective evaluations informed iterative refinements. Each round of feedback led to adjustments in the tuning of the model, helping it capture nuanced emotional expressions and improve the quality of its musical output.

This comprehensive evaluation and validation process ensures that the model not only meets technical standards but also resonates with listeners emotionally, making it suitable for applications where music generation must be both realistic and responsive to specific emotional inputs.

1. **Testing plan**

This testing plan outlines the evaluation and validation strategies to ensure that the Transformer-GAN model generates emotionally aligned and musically coherent compositions. The table below details each testing category, describing the method and expected outcomes for model performance, emotional accuracy, and system stability.

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Category** | **Test Description** | **Expected Outcome** | **Notes** |
| **Model Testing** |  |  |  |
| Validation Set Performance | Evaluate model on a separate validation set to monitor generalization and avoid overfitting. | Model should perform well on unseen data, showing good generalization and emotional alignment. | Use periodically during training |
| Emotion Accuracy | Measure how well the generated music aligns with intended emotional states (valence, arousal). | High correlation between the emotional content of the generated music and target valence/arousal values. | Use metrics like Mean Squared Error (MSE). |
| Human Evaluation | Collect feedback from listeners on emotional quality and coherence. | Listeners should perceive emotional accuracy and quality as intended by the model. | Survey listeners or use ratings on emotions. |
| **Unit and Integration Testing** |  |  |  |
| Unit Testing: Generator | Test the Generator to ensure it produces sequences with correct length and musical structure. | Generator should produce sequences that are coherent and structured, fitting the specified conditions. | Test for issues in sequence generation. |
| Unit Testing: Discriminator | Verify ability of the Discriminator to classify real vs. generated music accurately. | Discriminator should distinguish between real and generated sequences, improving Generator quality. | Test with controlled real/generated sequences. |
| Unit Testing: Gumbel-SoftMax | Ensure Gumbel-SoftMax handles discrete musical outputs reliably during training. | Gumbel-SoftMax should approximate sampling well without causing instability in generated outputs. | Monitor stability during training. |
| Integration Testing | Test entire system, ensuring Generator and Discriminator interact effectively in the adversarial loop. | Generator and Discriminator should work cohesively, producing music that aligns with emotional input. | Focus on interactions within GAN architecture. |

*8.1 Optional: Integration with User Interface (UI)*

If implemented, the UI allows users to interact with the system in real-time, adjusting emotional inputs and receiving immediate feedback in the form of dynamically generated music. The challenges in ensuring real-time data processing were addressed by optimizing the backend for speed and efficiency.

1. **Visualizing the music generation process**

*9.1 Sequence Diagram*תמונה שמכילה טקסט, צילום מסך, תרשים, גופן

התיאור נוצר באופן אוטומטי

Figure 7: Sequence diagram

תמונה שמכילה טקסט, צילום מסך, גופן, תרשים

התיאור נוצר באופן אוטומטי*9.2 Class Diagram*

Figure 8: Class diagram

1. **System Workflow**

תמונה שמכילה טקסט, צילום מסך, קבלה, קו

התיאור נוצר באופן אוטומטיThis diagram illustrates the step-by-step process flow for generating emotionally conditioned music using the Transformer-GAN architecture.

Figure 9: Transformer-GAN system workflow

*10.1 Detailed Process Flow*

*10.1.1 Data input*

Users specify the desired emotional states (e.g., valence and arousal), which are structured as numerical values to guide the music generation process.

*10.1.2 Data preprocessing*

As for the format, the emotional inputs are standardized, and other characteristics, such as the tempo and the key signature, are extracted. The next stage involves the application of encoding and scaling where necessary to the data collected. All these steps together make the inputs ready to be fed into the Transformer and GAN components in a way that allows for efficient and accurate generation of music.

*10.1.3 Model processing*

The Transformer component helps make the model attentive to long-range dependencies and also models the emotional restoration of continuity within the music sequence. At the same time, the GAN component produces and optimizes music sequences using the feedback of Discriminator which analyzes the realism of outputs and promotes enhancements for increasing realism. Altogether, all these components cooperate to create pieces with emotional impact and strong forms of organization .

*10.1.4 Post-processing*

Final touches are applied, like smoothing transitions and balancing volume, to align the output with the intended emotional tone.

*10.1.5 Output generation*

The final music piece is provided to the user. In real-time systems, users can adjust emotional inputs, dynamically updating the music to reflect changes.

1. **Software architecture overview**

The system comprises three main layers, each essential to the music generation process:

*11.1 Input Layer*

The function captures user-defined emotional states (valence and arousal), setting the emotional tone for music generation. This ensures that user input directly influences the output of the model, providing the context for emotionally aligned music.

*11.2 Processing Layer*

The transformer module is in charge of sequence modeling to address long-term dependencies of the musical data or the musical piece being sung and to generate music that goes well with the given emotions. Accompanying this the GAN module improves realism through an adversarial feedback loop. The Generator produces musical sequences, the Discriminator assesses and optimizes them to produce musically pleasing and emotionally affecting sequences.

*11.3 Output Layer*

תמונה שמכילה טקסט, צילום מסך, גופן, קו

התיאור נוצר באופן אוטומטיThe output layer aims to give the produced music to the user, guaranteeing that the output of the composed music is in harmony with the input of the user in terms of emotion and in addition to providing near real time listening experience. This process starts with the input layer receiving and storing emotional metrics and the next step is the processing layer receiving these inputs and analyzing them with the help of the Transformer that encapsulates the sequence context and the GAN that focuses on the musicality aspect. Finally, the output layer returns usable music to the user with appropriate emotions matching the given mood.

Figure 10: Output layer for generated music

The diagram visualizes the architecture, showing how data flows through the three main layers—from user input to final music generation. Each layer represents a step in the process, ensuring that the system works smoothly and produces emotionally resonant music in real-time.

1. **Next steps**

This project presents a structured plan to develop a Transformer-GAN model for emotion-based music generation, focusing on valence and arousal as emotional dimensions. By leveraging the sequence modeling strength of Transformers and the realism-enhancing capabilities of GANs, this system seeks to advance music generation technology.

The project builds insights from existing models like MusicLM and MuseNet, aiming to contribute to a unique approach. Through careful algorithm design, data strategy, and a phased development process, this model aspires to produce high-quality music that resonates emotionally with listeners.

*12.1 Next steps: Development plan for the coming semester*In the upcoming semester, the project will progress to the implementation phase, focusing on key milestones:

**Model Architecture setup**

* + Implement the Transformer and GAN components as defined in the architecture.
  + Achieve smooth integration between the sequence modeling of the Transformer and the adversarial processes of the GAN.

**Data collection and preprocessing**

* + Collect and refine datasets (e.g., AILABS17k, EMOPIA) to prepare for training.
  + Apply normalization, feature extraction, and data augmentation to optimize training inputs.

**Model training and optimization**

* + Conduct iterative training to improve emotional accuracy in music generation.
  + Employ Gumbel-Softmax and other methods to handle musical discreteness and stabilize training.

**Evaluation and validation**

* + Establish metrics to assess the quality and emotional alignment of generated music.
  + Perform both automated and human assessments to validate emotional resonance.

**Optional UI design**

Consider designing a user-friendly UI to allow users to select emotional inputs for music generation. This is a secondary goal and is not essential for project completion.

**Optional real-time feedback**

Real-time feedback may be added to the UI if resources allow, enhancing user interaction by adapting the music to input changes dynamically.

|  |  |  |  |
| --- | --- | --- | --- |
| **Testing Phase** | **Objectives** | **Testing Activities** | **Expected Outcomes** |
| **Model Performance Testing** | Assess the ability of the Transformer-GAN model to generate music that aligns with emotional states | - Evaluate accuracy of emotional output through metrics (valence/arousal correlation)  - Test music coherence | High emotional alignment in generated music, with coherence and consistent quality |
| **Unit Testing** | Verify each model component functions correctly | - Test the sequence outputs of the Transformer. - Validate GAN components (Generator and Discriminator). - Check emotion alignment module. | Each component should meet expected output standards, e.g., correct sequence lengths, accurate classification (real vs generated) |
| **System Integration Testing** | Ensure smooth collaboration between Transformer, GAN, and emotion alignment modules | - Test flow of data through model - Verify real-time adjustments based on emotion inputs | Data flows seamlessly through the system, with adaptable outputs that match intended emotional context |
| **Optional UI Testing** | Gather user feedback to refine user interface, if developed | - Collect feedback on ease of use, emotional accuracy perception - Assess real-time feedback effectiveness | UI should enhance user interaction without affecting core model functionality; improvements guided by feedback |
| **Final Validation** | Assess overall quality, emotional alignment, and coherence in generated music | - Use human listeners for subjective emotional accuracy ratings - Perform final checks on system stability and performance | Final output is musically coherent, emotionally resonant, and functionally robus |

*12.2 Testing and iteration plan for the upcoming semester*

The following table provides an overview of the testing and iteration plan for the upcoming semester. This plan focuses on optimizing the performance of the Transformer-GAN model and evaluating its ability to generate music aligned with specified emotional states. Optional UI testing is included if the interface is developed.

1. **References:**
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 2672–2680. <https://arxiv.org/abs/1406.2661>
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. <https://arxiv.org/abs/1810.04805>
4. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
5. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/1412.6980>
6. Neves, P. L. T., Fornari, J., & Florindo, J. B. (2022). Generating music with sentiment using transformer-GANs. arXiv preprint. <https://arxiv.org/pdf/2212.11134>
7. Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint. <https://arxiv.org/abs/1511.06434>
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008. <https://arxiv.org/abs/1706.03762>
9. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. Proceedings of the IEEE International Conference on Computer Vision, 5907–5915. <https://arxiv.org/abs/1612.03242>