Text-Conditioned Music Generation Using Variational Autoencoders and Transformer Networks

Mihir Birani

Abhiraj Mandal

***Abstract*—This project presents a system for generating music based on text descriptions of genres and moods. By combining variational autoencoders (VAEs) for learning musical structure with transformer-based text embeddings, our approach enables intuitive control over music generation through natural language descriptions.We utilise the MIDI file format to generate melodies using the VAE and we build a transformer network with Contrastive Learning that maps the user's prompt to a point in the VAE's latent space in order to generate music.**

I. Introduction

Music generation has been a long-standing challenge in artificial intelligence, with recent advances in deep learning enabling increasingly sophisticated approaches. While many existing systems can generate coherent musical pieces, providing intuitive control mechanisms remains challenging. Our work addresses this gap by enabling music generation conditioned on natural language descriptions of desired genres and moods. The MIDI output can then easily be transposed to any key or instrument of choice providing a great deal of flexibility to Music Producers as compared to models like OpenAI's Jukebox that work with raw audio.

We adapt a rather straightforward approach to generation using a Variational Autoencoder that learns a meaningful latent representation of musical structure from MIDI data, and we build a novel mapping approach which connects text to music, and allows users to generate music by simply describing genre and mood in natural language with an open dictionary.

II. Related Work

A. Music Generation with Deep Learning

Recent advances in music generation have leveraged various deep learning architectures. DeepBach used LSTMs for Bach-style chorales, while MuseNet applied transformers to multi-instrumental music generation. MusicVAE demonstrated the effectiveness of hierarchical VAEs for capturing long-term structure in music, inspiring our approach.

B. Text-to-Music Systems

Cross-modal generation between text and music is relatively unexplored compared to text-to-image synthesis. Jukebox pioneered raw audio generation conditioned on text but requires substantial computational resources. AIVA offers text-based controls but uses more restrictive parameterization. Our approach focuses on MIDI generation with intuitive text controls, balancing expressivity and computational efficiency.

C. Latent Space Alignment

Techniques for aligning latent spaces across modalities have been explored in text-to-image models like CLIP and DALL-E. We build on these concepts by aligning text embeddings with a musical latent space, adapting contrastive learning principles to the music domain. Contrastive Learning allows us to use a unified caption format, similar to CLIP while still allowing the model to learn differences among different inputs.

III. System Architecture

Our system consists of several key components that work together to enable text-conditioned music generation.

A. Overview

The architecture comprises two main subsystems:

- A music VAE that learns to encode and decode MIDI piano rolls
- A text-to-latent mapping network that connects natural language descriptions to the musical latent space

B. Music Representation and Processing

We represent music as piano rolls extracted from MIDI files, focusing on the melody instrument in each piece. Our preprocessing pipeline includes:

1) Instrument selection based on note density scoring
2) Transposition to a common key (C major/A minor)
3) Conversion to a binary piano roll representation
4) Arpeggiation of chords to create a monophonic representation
5) Segmentation into fixed-length sequences of 16 bars

C. Unified Caption Generation

The system accepts text descriptions of musical genres and moods, which are processed into a standardized format:

> Create a <genre1>, <genre2> song which is <mood1>, <mood2>, <mood3>...

These captions are then encoded using a pre-trained sentence transformer model (`all-MiniLM-L6-v2`) to create fixed-dimension embeddings.

D. Variational Autoencoder for Music

The core of our musical understanding is a VAE architecture with:

- A bidirectional LSTM encoder that processes piano roll sequences
- A latent space of 128 dimensions

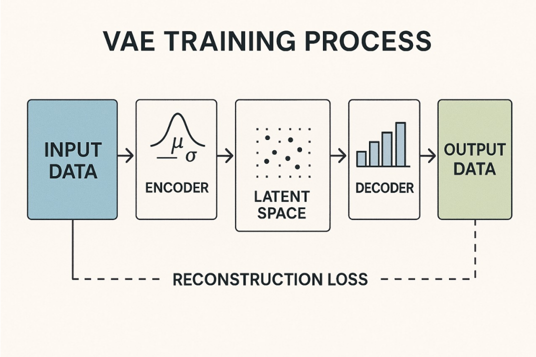

Fig. 1. VAE Training Procedure

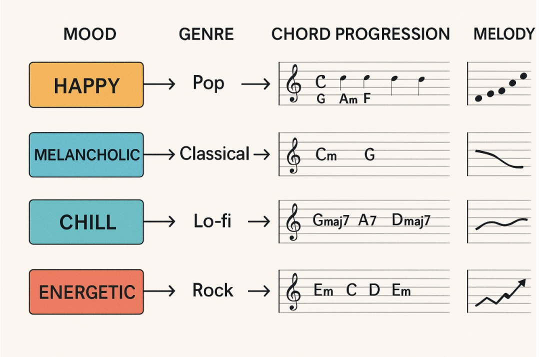

Fig. 2. Difference in Sound for different Genres and Moods

- A hierarchical decoder with:
  - A conductor LSTM that generates a sequence of embedding vectors
  - A note decoder LSTM that expands each embedding into a sequence of notes

This architecture enables the model to capture both local note patterns and longer-term musical structure.

E. Text-to-Latent Mapping

To connect text descriptions with musical latent vectors, we implemented a transformer-based mapping network that:

- Projects text embeddings through a transformer encoder
- Maps these embeddings to a latent vector
- Decodes the obtained latent vector via the VAE's decoder to give the final song

IV. Implementation Details

A. Music VAE Implementation

The music VAE implementation features several architectural refinements:

- Bidirectional LSTM encoder with 2 layers and 512 hidden units
- Latent space of 128 dimensions with normal prior
- Hierarchical decoder with conductor LSTM (2 layers) and note decoder (3 layers)
- Teacher forcing during training with scheduled sampling
- KL divergence annealing for the first 90 epochs
- Dropout rate of 0.1 for regularization

B. Text Embedding

For text embedding, we employed the pre-trained `all-MiniLM-L6-v2` model from the sentence-transformers library, which produces 384-dimensional embeddings. This model was chosen for its balance of performance and efficiency.

C. Text-to-Latent Transformer

The text-to-latent mapping network consists of:

- A transformer encoder with 3 layers and 8 attention heads
- Hidden dimension of 512
- Multi-layer perceptron for generating the sequence of latent vectors
- Positional embeddings added to maintain sequence ordering

D. Training Process

The system was trained in two stages:

1) **Music VAE training:**
   - Training on piano rolls extracted from the Lakh MIDI dataset
   - ELBO loss function with KL divergence annealing
2) **Text-to-latent mapping:**
   - Training on pairs of text captions and corresponding musical latent vectors
   - Contrastive NT-Xent Loss used which promotes similar pairs being together and dissimilar pairs being apart along with a temperature parameter to capture smoothness of the distribution.

E. Dataset Construction

The dataset for the text-to-latent mapping was constructed using:

1) Genre and mood tags from the MidiCaps dataset
2) MIDI files processed through our extraction pipeline
3) Text captions generated from the genre and mood combinations
4) Latent vectors obtained by encoding the processed MIDI files with the trained music VAE

V. Future Work

Several promising directions for future work include:

- In this work, our intermediate representation of music was one hot encoded,that is, we were allowing the music to only have a single note at a given time. This restricted us to rather simple compositions. We plan to build a new representation that eliminates this problem and can represent multiple notes at a given time in a data efficient manner

We also plan on:

- Expanding to multi-instrumental music generation

- Incorporating more specific musical attributes in the conditioning
- Developing interactive interfaces for real-time music generation with text guidance, where our model can directly be used to create the audio from different instruments and this interface can be linked to a DAW(digital Audio Workstation) to enable direct and seamless creation of songs

VI. Conclusion

We presented a system for text-conditioned music generation that combines the strengths of variational autoencoders for musical representation learning with transformer-based text-to-latent mapping. Our approach enables intuitive control over music generation through natural language descriptions of genres and moods. It demonstrates the effectiveness of our architecture in bridging the gap between textual descriptions and musical generation, opening new possibilities for creative applications in AI-assisted music composition.

References

- Hadjeres, G., Pachet, F., & Nielsen, F. (2017). DeepBach: a steerable model for Bach chorales generation. In International Conference on Machine Learning.
- OpenAI. (2019). MuseNet. Available at: https://openai.com/blog/musenet/
- Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. In International Conference on Machine Learning.
- Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341.
- AIVA Technologies. (2016). AIVA - Artificial Intelligence Virtual Artist. Available at: https://www.aiva.ai/
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. In International Conference on Machine Learning.