CHIME One Day Workshop 2023

The Open University, Milton Keynes
4 December 2023

Emotion-driven Motivic Development using Diffusion Models

Ronald K. Mo

School of Computer Science, University of Sunderland, United Kingdom,
ronald.mo@sunderland.ac.uk

Abstract—**Denoising diffusion probabilistic models, or *diffusion models*, have been successfully used for generating images, audio, and music. This work aims to investigate the potential of employing diffusion models to develop a motif composed by human composers to arouse specific emotions. To achieve this, a dataset consisting of *melodies* and their corresponding emotion label is constructed for training the diffusion model. The model is *conditioned* on the user-generated motif and a label displaying the desired emotion category, which opens up an opportunity for human composers to collaborate with computer technology in the field of music composition.**

I. Background

Diffusion models have showcased their capacity for generating realistic images [1]. In particular, a diffusion model comprises two processes. The *forward process* entails the repetitive addition of Gaussian noise to the input data x, such as images. Conversely, the *reverse process* is responsible for denoising a vector sampled from $p(z)$(i.e., the latent representation of x) in an iterative manner, ultimately restoring the input data to its original state. A well-trained diffusion model excels in learning the data distribution $p(x)$ within a provided set of data, enabling it to create novel data.

Beyond generating images, diffusion models have found applications in various GenAI tasks [3]. To facilitate the conditioning of the generated content, diffusion models often incorporate a *conditioning mechanism*. Conditional diffusion models are designed to learn the conditional distribution of $p(z \vee y)$ where y is the *conditioning input* such as class labels, text, or audio.

II. Overview

This work aims to investigate the capacity of employing diffusion models to expand upon motifs composed by human composers with the intention of bringing out specific emotions. To achieve this, a dataset is constructed, comprising melodies in symbolic music format paired with corresponding labels represent emotional categories. Emotion categories, instead of ratings, are used because they are more readily interpretable by composers. In the early stages of this research, the focus is placed on four primary emotional categories Happiness, Sadness, Calm, and Anger which correspond to the four quadrants of the valence-arousal plane [4].

A diffusion model is constructed using the previously mentioned dataset. To train this diffusion model, 16-bar melodies are initially *embedded* using a pre-trained 16-bar MusicVAE [5]. This embedding is then employed as input for the forward process of the diffusion model. During the reverse process, the vector sampled from the latent space, along with the conditioning inputs, are concatenated and denoised. The 2-bar motif, extracted from the first 2 bars of the melody, is embedded using a pre-trained 2-bar MusicVAE. The emotion label is encoded using simple one-hot embedding. During the inference phase, users are expected to provide a motif, along with an emotion label, to generate a full melody. The outcomes generated will be evaluated through both objective and subjective assessments.

III. Conclusion

This work presents a novel approach for crafting a human-composed motif that arouses specific emotions using diffusion models. This approach offers music composers the opportunity to interact with computers in their creative process.

References

[1] J. Ho et al., "Denoising diffusion probabilistic models," in *Advances in neural information processing systems 33*, 2020, pp. 6840–6851.

[2] M.W.Y. Lam et al., "BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis," in 2022-*10th International Conference on Learning Representations*, 2022.

[3] G. Mittal et al., "Symbolic Music Generation with Diffusion Models", in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021, pp. 468-475

[4] R. Mo et al., "The Effects of MP3 Compression on Perceived Emotional Characteristics in Musical Instruments," in *Journal of the Audio Engineering Society*, 2016, 64(11), pp. 858-867.

Ronald K. Mo is with the School of Computer Science. University of Sunderland, United Kingdom (corresponding author e-mail: ronald.mo@ sunderland.ac.uk).

[5] A. Roberts et al., “A hierarchical latent vector model for learning long-term structure in music,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 4364–4373.