CONTROLLED MUSIC GENERATION FROM UNLABELED DATA

Zhihao Ouyang, Keunwoo Choi
Bytedance
{zhihao.ouyang, keunwoo.choi}@bytedance.com

Yifei Li, Yuhong Zhang
Carnegie Mellon University
yifeili3@cs.cmu.edu

ABSTRACT

The Variational AutoEncoder (VAE) has demonstrated advantages in modeling for generation of sequential data with long-term structures such as symbolic music representations. In this paper, we propose a theoretically supported approach to enhance the interpretability of the VAE model by disentangling its latent space via the Gaussian Mixture Model (GMM) to generate highly controlled music through an unsupervised manner. Furthermore, with the Gaussian prior, latent variables retrieved from GMM clustering results are robuster than those of naive VAEs, especially in few-shot scenarios. With an implemented model, we observed the style of music is effectively controlled. It also reduced the amount of music data for music generation from hundreds per style to several. Demo link https://github.com/oyzh888/GMM_MusicVAE.

1. INTRODUCTION

Latest research works [1–3] have shown the effectiveness of VAE for encoding and decoding music sequences. The performance of VAE models largely depends on the quality of latent variables [4, 5]. With regard to music generation, latent variables encode crucial information including pitches, rhythms, dynamics, and textures for restoring a complete music piece. We hope to lucubrate the latent variables through an external model (GMM) to build a clear connection between the music style and latent vectors. A good example is beta-VAE [3, 6, 7], which provides a perspective to disentangle the music elements from latent variables. Comparing with similar work like MidiMe [8], our method is much lighter because we only need to perform clustering on latent space without any training procedure.

In our work, we choose GMM to further model the latent space because both GMM and VAE is consistent of holding the Gaussian prior. Firstly we train a VAE model on unlabeled data and use a GMM to cluster latent variables. Afterward, disentangled latent space is capable of generating music of specific class, moreover, it also enables us to manipulate given data for controllable music re-generation. For instance, we are able to interpolate or reconstruct them to novel musical pieces with combined styles of different composers or genres. Our pipeline is shown in Figure 1.

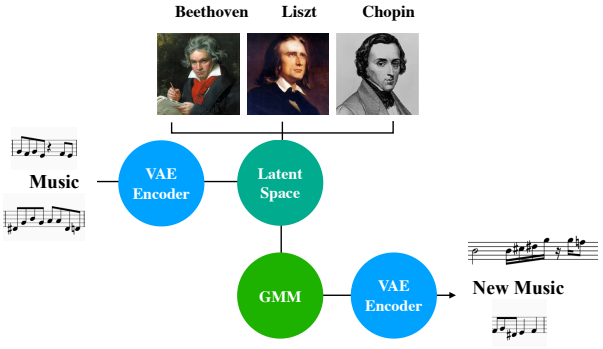

Figure 1. VAE+GMM pipeline for music generation. Firstly we extract latent variables from different composers' music pieces. Later, the GMM model clusters the latent variables and provides cluster membership information(eg. composer's style). We pass new latent variables sampled from GMM to VAE, then decode them to new music sequences.

To sum up, our contributions are as the following: 1) We propose a new method combining VAE and GMM to decompose the latent space of VAE model, which enable us to generate music of specific style unsupervisedly. We also provide related theoretical backgrounds. 2) We improve the robustness of the latent variable sampling, which improves the quality of the generated music, especially in few-shot scenarios. 3) We provide the two possible music applications of this model in supplementary web demo.

2. METHOD

Based on widely accepted MusicVAE models [1, 3, 6, 7], in our method, we introduce a GMM as a sub-module into the training-generating workflow. As shown in Figure 2, the whole workflow can be divided into two major steps: a training step and a generation step. Note that all the "music" we mention refers to music note sequences converted from midi files.

In the training step, we firstly use unlabeled music data to train a the encoder and decoder of a MusicVAE.

After the VAE is fully trained, we encode unlabeled music data into unlabeled latent vectors and use them to train a GMM. GMM will fit these data and cluster them into K

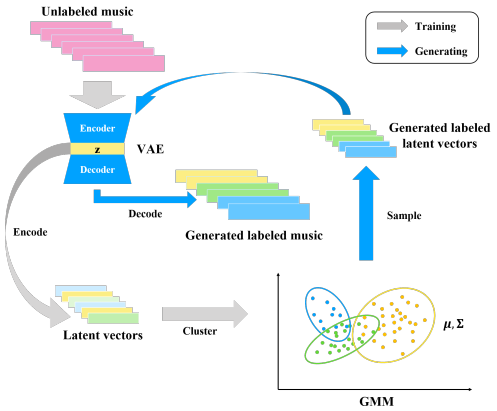

Figure 2. Overview of controlled music generation from unlabeled data with VAE and GMM. Both training procedure and generating procedure are included.

classes, where K implies the number of potential distinct styles the raw unlabeled data contains. Then GMM provides the parameters of the Gaussian distribution for each class.

In the generation step. Given a specific class, we sample latent vectors from the corresponding distribution and feed them to the decoder of the trained VAE. The decoder will decode them into real music belonging to the class we want.

2.1 Music encoding and latent vectors

We assume a VAE maps music data sample X to a latent vector z in the latent space. We also assume the latent vectors z follows a standard multivariate Gaussian distribution. Let θ be the parameters, and the VAE is trained to maximize the likelihood of X given θ:

$$\max_{\theta} P(X) = \int P(X|z;\theta)P(z)\mathrm{d}z, \tag{1}$$

where $P(X|z;\theta)$ is commonly assumed to be a Gaussian distribution [9].

To make this optimization feasible [9], the objective in equation 1 is derived as

$$\begin{aligned} &\log P(X) - \mathcal{D}[Q(z|X)||P(z|X)] = \\ &E_{z\sim Q}[\log P(X|z)] - \mathcal{D}[Q(z|X)||P(z)], \end{aligned} \tag{2}$$

where $\mathcal{D}(\cdot)$ is the KL-Divergence, and $Q(z)$ is PDF of the distribution of latent vectors while $Q(z|X)$ is the conditional probability given sample X. Similarly, we usually choose $Q(z|X)$ to be a Gaussian distribution $\mathcal{N}(z|\mu(X), \Sigma(X))$, where the mean $\mu(\cdot)$ and the variance $\Sigma(\cdot)$ are both learned through a neural network. Besides, the prior $P(z)$ is a standard multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Equation 2 implies one major theoretical basis of our method. Considering the RHS of equation 2, we can regard $P(X|z)$ as the decoder, and $Q(z|X)$ as the encoder [9]. Given a sample X, its corresponding latent vector follows a Gaussian distribution $Q(z|X)$. And any valid latent vector z sampled from the distribution $Q(z)$ can be decoded into a valid music sample by the decoder $P(X|z)$.

2.2 Unsupervisedly controlled generation with GMM

Given a fully trained VAE, we assume that every music sample X can be properly encoded into a latent vector z and we can regenerate a highly similar music sample of the corresponding X.

Now we have a batch of unlabeled music samples and encode them into latent vectors (by encoding and sampling as mentioned above). We know all the latent vector z's follows the distribution $Q(z)$. Our research goal is to classify these latent vectors into different classes according to the "message" they imply about the original music, like styles, composers, etc.

The proposed solution is based on Theorem 2.1:

Theorem 2.1 *Assuming the latent vector z of a sample X depends on the class label y of this sample, and that all the conditional probabilities $Q(z|y)$ are independent Gaussian distributions, then the latent space of a VAE model $Q(z)$ is a Gaussian mixture of these conditional probabilities:$Q(z) = \sum_y Q(z|y)P(y)$.*

Proof Let y be the class of music, and the conditional distribution of music sample X is $P(X|y)$. Consider $Q(z)$ again:

$$\begin{aligned} Q(z) &= \int_X Q(z|X)P(X)\mathrm{d}x \\ &= \sum_y \int_X Q(z|X)P(X|y)P(y)\mathrm{d}x \\ &= \sum_y Q(y,z) = \sum_y Q(z|y)P(y), \end{aligned} \tag{3}$$

where $Q(z|y)$ is the conditional distribution of latent vector z conditioned on the music class y.

Thus supposing each class of music is an independent component and there are K classes, we can decompose the distribution of the latent vectors as

$$Q(z) = \sum_{i=1}^{K} Q(z|y_i)P(y_i). \tag{4}$$

Then we use a GMM to describe the distribution of latent vectors. Given the unlabeled latent vectors, we assume they follow a Gaussian mixture distribution, and use a GMM to perform clustering. The number of classes K can be either pre-determined or found using the elbow method for clustering analysis.

After clustering, we get the individual latent vector distribution of each class of music. This performs a classification on the unlabeled data. Then for each class, we can sample from the corresponding distribution to get new latent vector z's, which follow the same distribution $Q(z)$ as the latent space of our VAE above. This means that using the decoder of the VAE, we can generate new music samples of this specific class. Hence with GMM our method, it is able to perform controlled music generation starting from unlabeled music data in an unsupervised way.

3. REFERENCES

[1] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," *arXiv preprint arXiv:1803.05428*, 2018.

[2] G. Hadjeres, F. Nielsen, and F. Pachet, "Glsr-vae: geodesic latent space regularization for variational autoencoder architectures," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017, pp. 1–7.

[3] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, "Midi-vae: Modeling dynamics and instrumentation of music with applications to style transfer," *arXiv preprint arXiv:1809.07600*, 2018.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013.

[5] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 3483–3491.

[6] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Advances in Neural Information Processing Systems*, 2018, pp. 2610–2620.

[7] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework." *International Conference on Learning Representations*, vol. 2, no. 5, p. 6, 2017.

[8] M. Dinculescu, J. Engel, and A. Roberts, Eds., *MidiMe: Personalizing a MusicVAE model with user data*, 2019.

[9] C. Doersch, "Tutorial on variational autoencoders," 2016.