Multimodal Variational Inference in Musical Style Transfer

Group Name: Music Box, **Students:** Shunshi Zhang (1003297238), Brian Cheong (1005061291), Katherine Williams (1002620256), KeJun Luo (1006459547)

Abstract

We propose to apply the gaussian mixture variational autoencoder (GMVAE) [1] models to musics of different genres to attempt style transfer. We will begin by training classifiers for each genre (for testing), followed by four GMVAE models that will be trained on separate sets of genres and evaluated according to pitch, instrument being played, the style, and note velocity (dynamics).

1 Summary of Technical Details

We can formulate style transfer as a dual probabilistic inference problem:

$$p(x) = \min_{\psi \in \mathcal{F}} \psi(x|z_{style}, z_{content}) p(z_{style}, z_{content}) \tag{1}$$

$$p(z_{style}, z_{content}) = \min_{\phi \in \mathcal{G}} \phi(z_{style}, z_{content}|x) p(x). \tag{2}$$

Here the two vectors $z_{style}, z_{content}$ represent the "style" (which represents features shared across an entire genre) and "content" (which are features localized to any particular sample); ϕ and ψ are the **encoder** and **decoder** represectively, and belong to parametric classes $\mathcal{F}, \mathcal{G}$. Variational autoencoders [2] jointly optimize this problem with the following loss function (through the evidence-lower bound):

$$L(x) = \log \psi(x|z) + D_{KL}(\phi(z|x), q(z)), \tag{3}$$

which can be directly minimized by function approximators. Here q represents a prior against which ϕ is to be regularized. Existing modelling [3] relies on end-to-end modelling of these probabilities using neural networks. While this is sufficient for applications which only require output samples, it is not interpretable for musicians interested in low-dimensional representations of music. Instead, we propose that both the transformation ϕ and the prior q be represented as a tractable multimodal distribution in $\mathbb{R}^d$, for instance as a sequential Gaussian Mixture model. In this case, z_{style} is a vector of length k, which represents:

$$p(x|z_{style}, z_{content}) = \sum_{i=1}^{k} \frac{\exp(z_{style}^{(i)})}{\sum_j \exp(z_{style}^{(j)})} \mathcal{N}(z_{content}|\mathbf{w}_i, \mathbf{\Sigma}_i), \tag{4}$$

i.e. softmax probabilities of belonging to a Gaussian. The parameters $\mathbf{w_i}, \mathbf{\Sigma}_i$ can be estimated using the Expectation-Maximization (EM) algorithm [4], and the encoder remains standard. For a sequential model, we simply require that ψ also be conditioned on a history of inputs, e.g. $\psi(x_t|h_t)$. This completes the motivation of our model, apart from some aspects relating to data engineering and cleaning.

2 Deliverables

The experiments will involve training several models, each being trained on different sets of genres ({Classical, Jazz, Pop }, {Bach and Mozart }). Each model should be capable of style transfer within these sets; these were chosen for their different levels of contrastability. A genre classifier will be trained to evaluate the style transfer performance metrics described below.

The main validation metrics for evaluating the project will be based on the metrics proposed in [3]. These include the change in classifier performance when evaluating bars before and after a transfer, and plotting the t-SNE graph of the learned latent vectors for the various genres. Qualitatively, the style transferred excerpts should keep the original melody while still being acceptable to listen to by general listeners. An intermediary metric while training the autoencoder is the reconstruction accuracy of the pitch, instrument being played, the style, and note velocity (dynamics).

3 Nice to Haves

Another approach we could try other than GMVAE for this topic are other multimodal sequential VAEs; if our results with GMVAE are not as promising as we expect we may try out alternative models.

In the conclusion of the article on MIDI-VAE [3], the authors mention extending the model, by integrating into hierarchical model, to be able to interpret and produce longer pieces of music as it's mentioned in their abstract that MIDI-VAE can only interpolate between short pieces. Though we likely will not have time in this project, it would be nice to try to integrate our GMVAE model to a hierarchical model and work with longer/larger pieces of music. In addition, since each model should be capable of two way style transfer, there is an opportunity to explore transferring to one style and then back to the original.

4 Review of related work

The article "MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer" [3] details the results of applying VAEs to MIDI files through inputting the note pitch, assignment instrument and velocity of the music, in the form of one hot vectors, and then training the model through the use of genre classifiers.

Our team realized that this model is constrained by only being able to process music in the form of MIDI files so we looked the article by Junxian He and company detailing the usage of a sequential VAE instead. [5]. However we realized that implementation of this model, while able to process audio data, uses speech rather than music and would not likely be applicable in our experiments and thus may need some adjustments to use.

The GMVAE found in the article by Yin-Jyun and company is more compatible [1]. The GMVAE formulated in this paper uses the latent variables timbre and pitch in instrumental music files. The authors have noted the applicability of the model not only in instrument note recordings but also in learning interpretable mixtures such as singer identity, music style. We intend to discover its potential in this study.

References

[1] Yin-Jyun Luo, Kat Agres, and Dorien Herremans. "Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders". In: *arXiv preprint arXiv:1906.08152* (2019).

[2] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[3] Gino Brunner, Andres Konrad, Yuyi Wang, et al. "MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer". In: *arXiv preprint arXiv:1809.07600* (2018).

[4] Todd K Moon. "The expectation-maximization algorithm". In: *IEEE Signal processing magazine* 13.6 (1996), pp. 47–60.

[5] Junxian He, Xinyi Wang, Graham Neubig, et al. "A probabilistic formulation of unsupervised text style transfer". In: *arXiv preprint arXiv:2002.03912* (2020).