Latent Space Regularization for Explicit Control of Musical Attributes

Ashis Pati [1] **Alexander Lerch** [1]

Abstract

Deep generative models for music are often restrictive since they do not allow users any meaningful control over the generated music. To address this issue, we propose a novel latent space regularization technique which is capable of structuring the latent space of a deep generative model by encoding musically meaningful attributes along specific dimensions of the latent space. This, in turn, can provide users with explicit control over these attributes during inference and thereby, help design intuitive musical interfaces to enhance creative workflows.

1. Introduction

In recent years, deep learning has emerged as the tool-of-choice for music generation models (Fiebrink et al., 2016; Briot & Pachet, 2018). While many of these deep generative models have been successfully applied to several different music generation tasks, e.g., monophonic music generation (Colombo et al., 2016; Sturm et al., 2016), polyphonic music generation (Boulanger-Lewandowski et al., 2012; Yang et al., 2017), creating musical renditions with expressive timing and dynamics (Huang et al., 2019; Oore et al., 2018), they are often found lacking in two critical aspects: control and interactivity (Briot & Pachet, 2018).

Latent representation-based models, such as Variational Auto-Encoders (VAE) (Kingma & Welling, 2014), have the potential to address this limitation as they are able to encode hidden attributes of the data (Carter & Nielsen, 2017). This is evident from properties such as attribute vectors (Mikolov et al., 2013; Roberts et al., 2018b) and semantic interpolations (Roberts et al., 2018a). Thus, improving the interpretability of latent spaces has been an active area of research. Methods to enforce semantic structure on the latent spaces have either used regularization methods (Lample et al., 2017; Hadjeres et al., 2017; Donahue et al., 2018), or transformation techniques (Engel et al., 2017; Adel et al., 2018). However, these have mostly been restricted to image generation tasks. The geodesic latent space regularization method proposed by Hadjeres et al. (2017) achieved some success for music data by encoding an attribute along a single dimension of the latent space. However, this method has not been tested for multiple attributes together and requires hyperparameter tuning for different attributes.

We propose a novel latent space regularization technique to improve the interpretability of latent spaces with respect to musically meaningful attributes understandable by humans. The proposed method can encode selected musical attributes along specific dimensions of the latent space. This enables the users to interactively control these attributes during inference time.

[1]Center for Music Technology, Georgia Institute of Technology, Atlanta, USA. Correspondence to: Ashis Pati <ashis.pati@gatech.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2. Method

The objective is to encode an attribute a along a dimension r of the latent space such that, as we traverse along r, the attribute value a of the generated music increases. For instance, if the attribute represents rhythmic complexity, sampling latent vectors with high values of r should result in music with high rhythmic complexity and vice versa. Mathematically, if $a_{\mathbf{x}_i} > a_{\mathbf{x}_j}$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two data-points generated using latent vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, then $z_i^r > z_j^r$ should hold for any arbitrary i and j. Here $\mathbf{z} : \left\{z^k\right\}, k \in [1, \mathbb{D}]$ is a vector in a $\mathbb{D}$-dimensional latent space.

This is accomplished by adding an attribute-specific regularization loss to the VAE training objective. To compute this loss, firstly, an attribute distance matrix $\mathcal{D}_a$ is computed for all examples in a training mini-batch: $\mathcal{D}_a(i,j) = a_{\mathbf{x}_i} - a_{\mathbf{x}_j}$, where $i, j \in [1, N]$, N is the number of examples in the mini-batch. Next, a similar distance matrix $\mathcal{D}_r$ is computed for the regularized dimension r of the latent vectors: $\mathcal{D}_r(i,j) = z_i^r - z_j^r$. The regularization loss is finally formulated as: $\mathcal{L}_{r,a} = \mathrm{MSE}(\tanh(\mathcal{D}_r) - \mathrm{sgn}(\mathcal{D}_a))$, where $\mathrm{MSE}(\cdot)$ is the mean square error, $\tanh(\cdot)$ is the hyperbolic tangent function, and $\mathrm{sgn}(\cdot)$ is the sign function. This formulation forces the values of the regularized dimension to have a monotonic relationship with attribute values while ensuring differentiability with respect to the latent vectors (and consequently the VAE-encoder parameters).

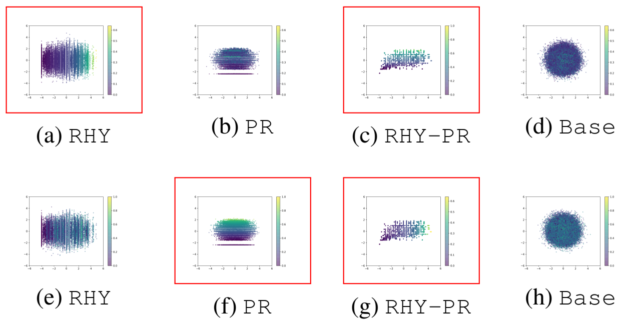

Figure 1. Attribute distribution for latent vectors obtained by encoding data from a held-out test set. The top row shows rhythmic complexity, bottom row shows pitch range. Sub-plots with a red border were regularized for the particular attribute. The x-axis denotes the value of the 1^{st} dimension while the y-axis denotes the value of the 3^{rd} dimension. Zoom in for higher resolution.

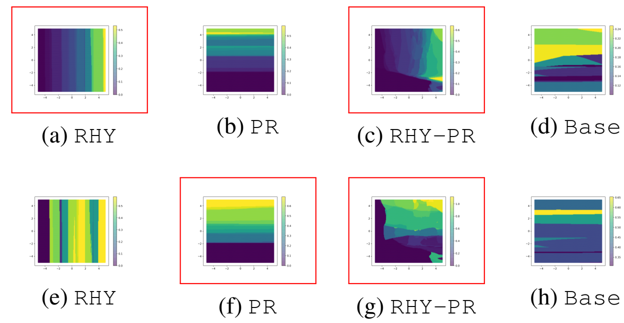

Figure 2. Attribute surface maps for decoded latent vectors on a 2-dimensional plane in the latent space (values for other dimensions are fixed). The arrangement of plots and the axis representation are similar to Fig. 1. Zoom in for higher resolution.

Figure 3. Measures generated by increasing the value of the regularized dimension (values of other dimensions are kept constant) for the RHY model. Rhythmic complexity increases gradually.

3. Experiments

Experiments were conducted using the proposed regularization technique for two attributes: *rhythmic complexity* and *pitch range*. For rhythmic complexity, Toussaint's metrical complexity measure was used (2002). This has been shown to correlate with human perception of rhythmic complexity (Thul & Toussaint, 2008). Pitch range was computed by taking the difference between the maximum and minimum MIDI pitch values of notes normalized by the range of notes.

Hierarchical VAE models (Roberts et al., 2018b) were trained on a dataset of monpphonic folk melodies in the symbolic domain (Sturm et al., 2016) to generate single measures of music. Models RHY and PR were trained with rhythmic complexity regularized along the 1^{st} dimension and pitch range regularized along the 3^{rd} dimension, respectively. A third model RHY-PR was trained which jointly regularized both attributes along these dimensions. For comparison, a fourth model Base was trained with no regularization. Other training parameters (e.g., optimizer, learning rate, batch-size etc.) were kept consistent across the three models.

All models achieved a low NLL loss (≈ 0.003) with high reconstruction accuracy ($\approx 99\%$) on a held-out test set. However, the attribute distributions of the latent vectors (obtained by passing data from the test set through the VAE-encoder) are very different (see Fig. 1). There is a clear ordering of the attributes along the respective regularized dimensions for the regularized models while there is no such structure for the Base model.

Attribute surface maps obtained by decoding latent vectors on a 2-dimensional plane (comprised of the regularized dimensions) of the latent space also show a similar structure (see Fig. 2). The attribute values are monotonically ordered along the corresponding regularized dimensions. Moving along these dimensions also produces measures with increasing value of the corresponding attribute (see Fig. 3).

Models were also evaluated using the interpretability metric (Adel et al., 2018). This was modified slightly by replacing the linear classifier with a linear regression model. The regression scores (higher is better) are: RHY: 0.84 (rhythmic complexity), PR: 0.96 (pitch range), RHY-PR: 0.90 (average). In contrast, the Base model only manages 7.9e−06 (average). More information is available online.[1]

4. Conclusion

The results demonstrate that the proposed method is able to encode selected musical attributes along different dimensions of the latent space. This has potential to provide users with more intuitive control over the generated music. The regularization loss is simple to compute (as long as the attribute values can be computed) and requires no hyperparameter tuning. Future work will involve carrying a more thorough evaluation (using objective and subjective methods) by comparison with other latent space regularization methods (Hadjeres et al., 2017; Lample et al., 2017).

[1] https://github.com/ashispati/AttributeModelling

References

Adel, T., Ghahramani, Z., and Weller, A. Discovering interpretable representations for both deep generative and discriminative models. In *Proc. of the 35th International Conference on Machine Learning*, pp. 50–59, Stockholmsmässan, Stockholm, Sweden, 2018.

Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proc. of 29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, 2012.

Briot, J.-P. and Pachet, F. Deep learning for music generation: Challenges and directions. *Neural Computing and Applications*, Oct 2018. ISSN 1433-3058. doi: 10.1007/s00521-018-3813-6.

Carter, S. and Nielsen, M. Using artificial intelligence to augment human intelligence. *Distill*, 2017. doi: 10.23915/distill.00009. https://distill.pub/2017/aia.

Colombo, F., Muscinelli, S., Seeholzer, A., Brea, J., and Gerstner, W. Algorithmic composition of melodies with deep recurrent neural networks. In *Proc. of the 1st Conference on Computer Simulation of Musical Creativity (CSMC)*, 2016.

Donahue, C., Lipton, Z. C., Balsubramani, A., and McAuley, J. Semantically decomposing the latent spaces of generative adversarial networks. In *Proc. of International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

Engel, J., Hoffman, M., and Roberts, A. Latent constraints: Learning to generate conditionally from unconditional generative models. In *Proc. of International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

Fiebrink, R., Caramiaux, B., Dean, R., and McLean, A. *The machine learning algorithm as creative musical tool*. Oxford University Press, 2016.

Hadjeres, G., Nielsen, F., and Pachet, F. GLSR-VAE: Geodesic latent space regularization for variational autoencoder architectures. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pp. 1–7. IEEE, 2017.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., and Eck, D. Music transformer. In *Proc. of International Conference of Learning Representations (ICLR)*, New Orleans, USA, 2019.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *Proc. of International Conference of Learning Representations (ICLR)*, Banff, Canada, 2014.

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5967–5976, 2017.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3111–3119, 2013.

Oore, S., Simon, I., Dieleman, S., Eck, D., and Simonyan, K. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, pp. 1–13, 2018.

Roberts, A., Engel, J., Oore, S., and Eck, D. Learning latent representations of music to generate interactive musical palettes. In *IUI Workshops*, 2018a.

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., and Eck, D. A hierarchical hatent vector model for learning long-term structure in music. In *Proc. of the 35th International Conference on Machine Learning (ICML)*, pp. 4364–4373, Stockholmsmässan, Stockholm, Sweden, 2018b.

Sturm, B. L., Santos, J. F., Ben-Tal, O., and Korshunova, I. Music transcription modelling and composition using deep learning. In *Proc. of the 1st Conference on Computer Simulation of Musical Creativity (CSMC)*, Huddersfield, UK, 2016.

Thul, E. and Toussaint, G. T. On the relation between rhythm complexity measures and human rhythmic performance. In *Proc. of the C3S2E Conference*, pp. 199–204, Montreal, Quebec, Canada, 2008.

Toussaint, G. T. A mathematical analysis of African, Brazilian, and Cuban clave rhythms. In *Proc. of BRIDGES: Mathematical Connections in Art, Music and Science*, pp. 157–168, 2002.

Yang, L.-C., Chou, S.-Y., and Yang, Y.-H. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. In *Proc. of International Society of Music Information Retrieval Conference (ISMIR)*, pp. 324–331, Suzhou, China, 2017.