EVALUATION OF LATENT SPACE DISENTANGLEMENT IN THE PRESENCE OF INTERDEPENDENT ATTRIBUTES

Karn N. Watcharasupat[1,2] **Alexander Lerch**[1]
[1]Center for Music Technology, Georgia Institute of Technology, Atlanta, GA, USA
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
karn001@e.ntu.edu.sg, alexander.lerch@gatech.edu

ABSTRACT

Controllable music generation with deep generative models has become increasingly reliant on disentanglement learning techniques. However, current disentanglement metrics, such as mutual information gap (MIG), are often inadequate and misleading when used for evaluating latent representations in the presence of interdependent semantic attributes often encountered in real-world music datasets. In this work, we propose a dependency-aware information metric as a drop-in replacement for MIG that accounts for the inherent relationship between semantic attributes.

1. INTRODUCTION

Disentanglement learning has been an influential field of studies for controllable music generation with variational autoencoders (VAEs). A number of previous studies have attempted supervised disentanglement learning techniques on several semantic attributes such as rhythm, pitch range [1], note density, contour [2], arousal [3], style [4], and genre [5] to varying degrees of success. However, learning to simultaneously manipulate multiple attributes, in particular, remains a difficult task to both achieve and objectively evaluate [6] due to the limitation of current metrics.

One major issue with popular disentanglement metrics [7], such as mutual information gap (MIG) [8], separate attribute predictability (SAP) [9], and modularity [10], is that they were designed for independent generative factors, rather than real-world semantic attributes. As semantic attributes related to music are often highly interdependent, these metrics do not provide an accurate reflection of the 'quality' of learnt latent representation regularized for multiple interdependent attributes. Information inherently shared between attributes is penalized in the same way as that due to undesired entanglement issues.

In this work, we propose a dependency-aware metric based on mutual information (MI) to act as a drop-in replacement for MIG. Preliminary experiments were carried out to demonstrate the benefits of the proposed metrics over MIG.

© K. N. Watcharasupat, and A. Lerch . Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** K. N. Watcharasupat, and A. Lerch , "Evaluation of Latent Space Disentanglement in the Presence of Interdependent Attributes", in *Extended Abstracts for the Late-Breaking Demo Session of the 22nd Int. Society for Music Information Retrieval Conf.,* Online, 2021.

2. PROPOSED METRICS

Consider a set of attributes $\{a_i\}_{i=1}^{M}$ and a latent vector $\boldsymbol{z} \in \mathbb{R}^D$ with $M \leq D$. Without loss of generality, for $i \leq M$, we assume z_i is regularized for a_i. The remaining dimensions are unregularized. $\mathcal{H}(\cdot)$ denotes entropy while $\mathcal{I}(\cdot,\cdot)$ denotes mutual information.

MIG was proposed in [8] to measure the degree of disentanglement in a latent space. The idea behind MIG can be said to measure: *for each attribute, the normalized difference between the mutual information between the attribute and its most informative latent dimension, and that between the attribute the second-most informative latent dimension*. Mathematically, MIG is given by

$$\text{MIG}(a_i) = \left(\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j)\right) / \mathcal{H}(a_i), \qquad (1)$$

where $j = \arg\max_{k \neq i} \mathcal{I}(a_i, z_k)$. It is reasonable to assume $i = \arg\max_k \mathcal{I}(a_i, z_k)$ in a supervised setting; otherwise MIG takes negative values to indicate regularization failure. The normalization is given by $\mathcal{H}(a_i)$, which would be the maximum possible difference in MI between a latent dimension z_i coding perfectly for a_i, i.e., $\mathcal{I}(a_i, z_i) = \mathcal{H}(a_i)$ and the second-most containing no information about a_i, i.e., $\mathcal{I}(a_i, z_j) = 0$. As such, MIG is bounded above by one.

However, given the interdependence of semantic attributes, if $j \leq M$, the ideal value of the difference $\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j)$ is no longer $\mathcal{H}(a_i)$ since

$$\mathcal{I}(z_j, a_j) > 0 \wedge \mathcal{I}(a_i, a_j) > 0 \implies \mathcal{I}(a_i, z_j) > 0. \quad (2)$$

For regularized latent dimensions, we consider a pair of inherently entangled attributes (a_i, a_j), i.e., $\mathcal{I}(a_i, a_j) > 0$. Under the ideal case where z_i is fully informative [7] about a_i, i.e., $\mathcal{H}(a_i|z_i) = 0$, we have

$$\begin{aligned} &\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j) \\ &\quad = [\mathcal{H}(a_i) - \mathcal{H}(a_i|z_i)] - [\mathcal{H}(a_i) - \mathcal{H}(a_i|z_j)] \quad (3) \\ &\quad = \mathcal{H}(a_i|z_j) \qquad \because \mathcal{H}(a_i|z_i) = 0. \quad (4) \end{aligned}$$

Moreover, in the ideal case, z_j and a_j also have an invertible mapping between each other, this means that

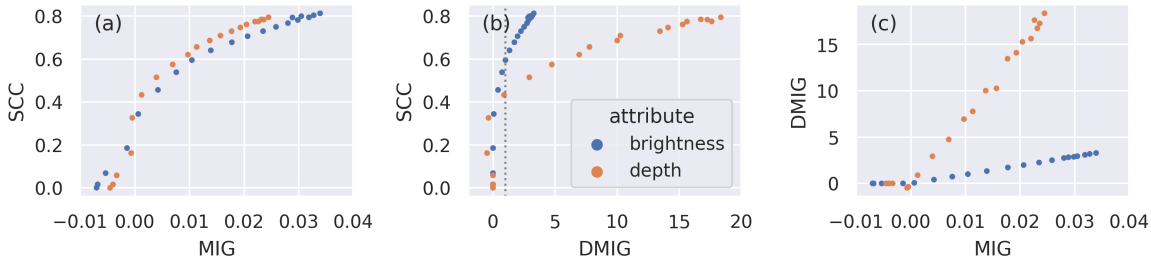

Figure 1: Plots of (a) SCC against MIG, (b) SCC against DMIG, and (c) DMIG against MIG, on the validation set.

$\mathcal{H}(a_i|z_j) = \mathcal{H}(a_i|a_j)$. Hence, in the ideal case, the difference is given by

$$\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j) = \mathcal{H}(a_i|a_j). \tag{5}$$

As such, we extend the definition of mutual information gap to the dependency-aware mutual information gap (DMIG) as follows

$$\text{DMIG}(a_i) = \begin{cases} \left(\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j)\right) / \mathcal{H}(a_i|a_j) & j \leq M \\ \left(\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j)\right) / \mathcal{H}(a_i) & j > M, \end{cases} \tag{6}$$

where $j = \arg\max_{k \neq i} \mathcal{I}(a_i, z_k)$. DMIG remains faithful to the core idea of MIG but modifies the normalization to properly account for inter-attribute dependencies. When a_i and a_j are independent, $\mathcal{H}(a_i) - \mathcal{I}(a_i, a_j) = \mathcal{H}(a_i)$ and the DMIG reduces to vanilla MIG.

Note that in the case of continuous random variables, differential entropy can be negative, unlike discrete Shannon entropy. This is particularly evident with conditional differential entropy and may result in DMIG values above unity whenever $\mathcal{H}(a_i|z_i)/\mathcal{H}(a_i|z_j)$ is negative.

3. EXPERIMENTS

To illustrate the key features of the dependency-aware metrics, we evaluate the latent space of a VAE model trained to reconstruct raw musical audio while being regularized for two highly correlated attributes [1] .

3.1 Data and model

We use the NSynth dataset [11], which is a large-scale dataset of musical notes played by various instruments with diverse timbral qualities. The dataset provides 4-second snippets sampled at 16 kHz. From the raw audio provided by NSynth, we extract two semantic attributes, namely, *brightness* and *depth* using the AudioCommons Timbral Model [12]. Since both the brightness and depth features are heavily influenced by the spectral distribution of the sound [13], they are strongly correlated.

We trained a convolutional VAE model to reconstruct the log-magnitude spectrogram of the audio and obtain reconstructed time-domain audio using a phase-bypass reconstruction. The models are trained using the attribute-regularized β-VAE loss function [2, 14, 15]

$$\mathcal{L} = \mathcal{R}(\hat{\mathbf{x}}; \mathbf{x}) + \beta \mathcal{D}(\mathbf{z}) + \gamma \textstyle\sum_i \mathcal{A}(z_i; a_i), \tag{7}$$

where $\mathcal{R}(\cdot)$ is the reconstruction loss implemented via the mean square error on the log-magnitude spectrograms, $\mathcal{D}(\cdot)$ is the KL divergence term with a standard normal prior, and $\mathcal{A}(\cdot)$ is the AR-VAE regularization from [2]. We used $D = 512$, $\beta = 1$, and $\gamma = 10$.

3.2 Results

Figure 1 plots the MIG, DMIG, and Spearman correlation coefficient (SCC) of the attributes (brightness and depth) with respect to their respective regularized latent dimensions on the validation set over the course of the training. Due to the high correlation between brightness and depth, for most of the training, the most and second-most informative latent dimensions in MIG/DMIG are the regularized ones that encode for the attributes.

As seen from Figure 1(a), the MIG values are generally very low (in the order of 10^{-2}, out of maximum 1) despite the SCC indicating successful encoding of the attribute information into the latent dimension. This is due to the high mutual information between brightness and depth, resulting in a very low true bound for MIG. On the other hand, we can observe from Figure 1(b) that DMIG reflects more clearly the quality of the latent space as it encodes the attribute; the rapid improvement in SCC mostly occurred before DMIG reaches one (dotted line). In Figure 1(c), the highly linear relationship between MIG and DMIG further demonstrates the idea that DMIG is simply MIG renormalized to better reflect the dependencies between semantic attributes coded by the model. Admittedly, the peculiarities of differential conditional entropy and the practical computation of mutual information and entropy estimates [16] have contributed to a DMIG range that is much larger than vanilla MIG. We will be working to resolve this limitation in future work.

4. CONCLUSION

In this work, we propose a dependency-aware extension to a popular disentanglement metrics, mutual information gap (MIG), to better account for inter-attribute dependencies often observed in real-world datasets. Key features of the proposed dependency-aware MIG were demonstrated via an experiment on an audio dataset with highly correlated timbral attributes.

[1] See the supplementary materials for full experimental details at https://github.com/karnwatcharasupat/dependency-aware-mi-metrics.

5. ACKNOWLEDGEMENT

K. N. Watcharasupat acknowledges the support from the CN Yang Scholars Programme, Nanyang Technological University, Singapore.

6. REFERENCES

[1] A. Pati and A. Lerch, "Latent Space Regularization for Explicit Control of Musical Attributes," in *Extended Abstracts for ML4MD, ICML*, 2019.

[2] ——, "Attribute-based regularization of latent spaces for variational auto-encoders," *Neural Comput. Appl.*, vol. 33, no. 9, pp. 4429–4444, 2020.

[3] H. H. Tan and D. Herremans, "Music fadernets: Controllable music generation based on High-Level features via Low-Level feature modelling," in *Proc. ISMIR*, 2020.

[4] Y. N. Hung, I. T. Chiang, Y. A. Chen, and Y. H. Yang, "Musical composition style transfer via disentangled timbre representations," in *Proc. IJCAI*, 2019, pp. 4697–4703.

[5] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, "MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer," *Proc. ISMIR*, pp. 747–754, 2018.

[6] A. Pati and A. Lerch, "Is Disentanglement enough? On Latent Representations for Controllable Music Generation," in *Proc. ISMIR*, 2021.

[7] K. Do and T. Tran, "Theory and Evaluation Metrics for Learning Disentangled Representations," in *Proc. ICLR*, 2020.

[8] T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Proc. NeurIPS*, 2018.

[9] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," in *Proc. ICLR*, 2018.

[10] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the F-statistic loss," in *Proc. NeurIPS*, 2018, pp. 185–194.

[11] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proc. ICML*, 2017, pp. 1771–1780.

[12] A. Pearce, S. Safavi, T. Brookes, R. Mason, W. Wang, and M. Plumbley, "Deliverable D5.8 - Release of timbral characterisation tools for semantically annotating non-musical content," Audio Commons Initiative, Tech. Rep., 2019.

[13] ——, "Deliverable D5.2 - First prototype of timbral characterisation tools for semantically annotating non-musical content," Audio Commons Initiative, Tech. Rep., 2017.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. ICLR*, 2014.

[15] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2017.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.