DO UNCONDITIONAL DEEP GENERATIVE MODELS SPONTANEOUSLY LEARN HOW TO ENCODE HUMAN-INTERPRETABLE MUSICAL ATTRIBUTES?

Colombo Marco Furio, Malaman Vittoria, Pettenò Matteo

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20122 Milano, Italy
[marcofurio.colombo,vittoria.malaman,matteo.petteno]@mail.polimi.it

ABSTRACT

Deep generative models have proved to be powerful creative devices, suitable for plenty of artistic applications. A leading role is played by variational autoencoders: neural network architectures trained to map random inputs in their latent space to new samples, representative of the learned data. However, due to its high dimensionality and the non-linearity of the generation process, the structure of the latent space is hard to grasp, thus limiting the usefulness of these models. It is with this mind that we develop the proposed analysis. We examine a 2-bar pre-trained unconditional model of MusicVAE, a recurrent variational autoencoder by Magenta. Our study is focused on investigating the presence of any correlations between the topological structure of the latent space and different human-interpretable musical attributes of the output. To allow an organized mapping of the latent space, we adopt a Latin Hypercube Sampling, generating controlled random samples. Our analysis aims at depicting any intrinsic correlation between the latent space structure and the characteristics of the output, supplying a valid starting point for controllable musical sequence generation via latent space navigation, without the need of further conditioning techniques.

Index Terms— Deep generative models, Variational autoencoders, Latent space topological structure, musical attributes, Latin Hypercube Sampling

1. INTRODUCTION

Recent breakthroughs in deep generative systems have enabled the creation of high fidelity media such as image, speech, text, and music. These developments are becoming more and more relevant also for creative domains: trained on a large data corpus, such generative models are increasingly being used to synthesize new original content. Among these, variational autoencoders (VAEs) stand out: the continuity of their latent space enables interpolation, which provides a creative way to explore artistic ideas [1]. One of the main challenges associated with the attainment of such goals regards the user control over the generated output. This is strictly related to the mechanisms underlying neural network architectures, which support conditional and controlled generation [2]. Many researchers have recently focused on the implementation of conditional VAEs, operating both upstream and downstream of the training phase. In this regard, a popular approach consists of developing user-specified constraints during training, either by training on a curated subset of data or with conditioning variables. Another important technique involves post-hoc learning latent constraints and enables conditional generation without retraining the model, by applying behavioral constraints on an embedding space [3]. The proposed work fits into this context; in particular, it focuses on the structural analysis of the latent space of a trained unconditional deep generative model studying its topological structure and its role in the inference process. Inquiring about the navigability of the latent space entails the hypothesis of a spontaneous organization of the embedding, which may overcome the necessity of additional conditioning. For our analysis, we worked on a 2-bar pre-trained model of MusicVAE by Magenta and focused on spotting any correlations between the topological structure of the latent space and different human-interpretable musical attributes. From now on we'll just refer to these as attributes. To detect such correlations, we perform a regularization of the latent space, using a Latin Hypercube Sampling (LHS); then, we investigate the trend of such attributes in relation to the sample's coordinates within the embedding.

The following article is organized as follows: Section 2 provides a description of architecture of the pre-trained model we work on, preceded by a brief overview of VAEs. Section 3 discusses the proposed methodology for the latent space mapping and visualization and outlines the main phases of our analysis. Section 4 goes into detail of the implementation and the examined musical attributes. It also presents the results and their implications, corroborated by a statistical test. Section 5 draws conclusions on the proposed study.

2. BACKGROUND ON VAE

VAEs are latent space models capable of learning the fundamental characteristics of a training dataset, i.e. translating the variation of real data into a lower-dimensional space, referred to as the latent space. The core of a VAE is the combination between encoder and decoder: the former, given a datapoint $\mathbf{x}$, produces an approximate posterior distribution $q_\theta(\mathbf{z}|\mathbf{x})$ over the possible values of a latent variable $\mathbf{z}$ from which the datapoint $\mathbf{x}$ could have been generated; the latter, given a code $\mathbf{z}$ produces a distribution $p_\phi(\mathbf{x}|\mathbf{z})$ over the possible corresponding values of $\mathbf{x}$. As for θ and ϕ, they represent the weights which parametrize $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$ in the neural networks [4, 5].

The distinctive feature of VAEs is the nature of $\mathbf{z}$, obtained by the dataset compression: instead of directly learning the latent variables, VAEs learn a mean μ and a standard deviation σ which parametrize a probability distribution for each of these latent variables. One of the main issues of this architecture is that we cannot backpropagate gradients through a sampling layer due to its stochastic nature, since this operation requires deterministic nodes. To address this problem the reparametrization trick is adopted; it consists of recasting the sampled latent vector z as the sum of a fixed μ and σ vectors, as in (1)

$$\mathbf{z} = \mu + \sigma \odot \epsilon \tag{1}$$

where $\odot$ indicates the element-wise product.
MusicVAE provides several implementations of the previously described high-level architecture, according to the selected configuration. In our analysis, we worked on a pre-trained 2-bar model, realized exploiting the combination of a Long Short Term Memory (LSTM) encoder with a flat LSTM decoder. The idea at the basis of the functioning of the bidirectional encoder is to split the state neurons into a part that is responsible for the positive time direction (forward states) and a part for the negative time direction (backward states); outputs from forward states are not connected to inputs of backward states, and vice versa. The final state vectors are then concatenated and fed into two fully-connected layers to produce the latent distribution parameters μ and σ. The two-layer bidirectional LSTM network state size is 2048 for all layers and produces an embedding having 512 latent dimensions [6, 7]. As regards the decoder, it is composed of a two-layer RNN composed of 256 LSTM cells, producing a single categorical output [8].

3. METHODOLOGY

In the following section, we propose a technique for the analysis of the latent space of a trained deep generative model.
The adopted technique employs a regular and replicable latent space mapping framework, which substitutes the random Gaussian sampling carried out by MusicVAE in inference mode. Since we are handling a highly multi-dimensional space, we need to choose a suitable sampling method: techniques traditionally employed in two or three dimensions do not scale well enough. LHS generates a near-random sampling from a multidimensional distribution and simultaneously stratifies them on all input dimensions to improve the input space coverage [9]. This technique fits our needs because it allows a well organized mapping of the hyperspace, with a feasible number of sample. We associate the set of points produced by LHS to the centroids of different regions in the hyperspace, then sample from Gaussian distributions centered in each of those points and use the collected embeddings as input for the model's decoder, obtaining MIDI files as output. From these we can proceed with the evaluation of a set of attributes and check for the presence of any intrinsic correlation patterns.
In order to better grasp the topology of the latent space, we also produce a bidimensional visualization. Achieving this from a high-dimensional space is no trivial task, since we need to maintain the properties of our organized sampling through the transformation. We implement a Multi Dimensional Scaling (MDS), which performs the needed dimensionality reduction, while keeping the original high-dimensional Euclidean distances. [10].
A further step in our analysis consists of an additional statistical investigation of the three regions associated with the highest, mid, and lowest mean attributes values: the attributes' distribution of the samples obtained from each region is compared, both visually and statistically.

4. EVALUATION

In this section, we discuss the details of our analysis and evaluate the obtained results.
To carry out our study, we extend the 2-bar pre-trained MusicVAE model. The latter has a D-dimensional latent space, with $D = 512$. We perform LHS and obtain a total of 100 points $g_i,\ i = 1, \dots, G$, collected in the vector $\mathbf{g}$ and representing the centroids of the hyperspace regions. For each g_i, we compute $X = 64$ multivariate Gaussian distribution $\mathcal{N}_D(\mu,\, \sigma^2)$, having $\mu = g_i$ and the value of σ is chosen so to have a probability of 95% of producing samples within half the minimum distance between two centroids from the Gaussian sampling, meaning that $3 \cdot \sigma = m/2$, where m is the minimum euclidean distance between two centroids. This ensures a smooth sampling mostly from within the region, with a small chance of sampling from nearby regions.
We obtain a total of $S = X \cdot G$ samples, whose coordinates are stored in the latent codes array:

$$\mathbf{x} = \big[[x_{1,1}\ \dots\ x_{X,1}]\ \dots [x_{1,G}\ \dots\ x_{X,G}]\big]\,.$$

Each element of the vector $\mathbf{x}$ is a batch that we give as input to MusicVAE's decoder, which in turns generates a total of S MIDI files.
We list below the attributes we choose for the output's analysis. These are defined for M music measures, composed of N symbols $\{m_t\},\ t \in [0, N)$ [11].

- **Toussaint's metrical complexity measure**: It allows the computation of the rhythmic complexity of a given measure by assigning weights to different metrical locations: low weights are associated with on-beat events, while high ones correspond to off-beat locations. The weights are then collected into a complexity coefficient array f and the attribute is computed by taking a weighted average of the note onset locations with f [12][13].
- **Pitch range**: Corresponds to the normalized difference between the maximum and minimum MIDI pitch values:

$$p(M) = \frac{1}{R}[\max_{t\in[0,N)}(\text{MIDI}(m_t)) - \min_{t\in[0,N)}(\text{MIDI}(m_t))], \quad (2)$$

- **Note density**: Counts the number of notes per measure normalized by the total length of the measure sequence:

$$d(M) = \frac{1}{N}\sum_{i=0}^{N-1} \text{ONSET}\,(m_t)\,, \quad (3)$$

- **Contour**: Measures the degree to which the melody moves up or down; it is computed by summing up the difference in pitch values of all the notes in the measure:

$$c(M) = \frac{1}{R}\sum_{t=0}^{N-2}\left[\text{MIDI}\,(m_{t+1}) - \text{MIDI}\,(m_t)\right], \quad (4)$$

where MIDI $(\cdot)$ computes the pitch value in MIDI for the note; ONSET $(\cdot)$ is 1 if its argument is a note onset and 0 otherwise; the normalization factor R depends on the range of the dataset: since we're considering piano MIDI files, $R = 88$.
The results of each batch are collected into four vectors, with the following structure

$$\mathbf{a} = \big[[a_{1,1}\ \dots\ a_{X,1}]\ \dots [a_{1,G}\ \dots\ a_{X,G}]\big]\,.$$

4.1. Results and discussion

A first approach to evaluate the presence of an intrinsic organization within the latent space is to depict any correlation between the chosen attributes: to do so, we compute the Pearson correlation coefficient ρ, as reported in Fig. 1. Note density and contour show a strong correlation ($\rho = 0.87$): a higher number of notes tends to be

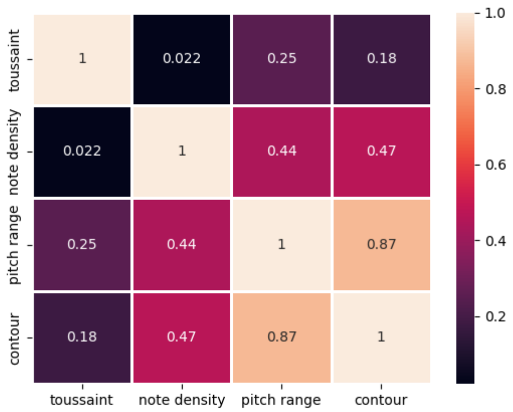


Figure 1: Pearson correlation between attributes.

spread on a wider pitch domain so as to increase the contour.

Next we look for any linear correlation between the output's characteristics and any of the dimensions of the latent space. This is done by performing two different tests: we compute ρ element-wise between $\mathbf{x}$ and $\mathbf{a}$; then, we repeat the same operation between $\mathbf{g}$ and $\bar{\mathbf{a}}$, i.e., the vector containing the mean value of the measured attributes, associated to each region of the hyperspace centered in g_i. The values of ρ suggest the presence of an intrinsic directionality within the latent space which influences the output characteristics only in terms of note density. This trend is clearly displayed in Fig. 2; the depicted linear trend confirms the relation between a component of the embedding which is topologically structured in accordance with the inspected qualities of the model's outputs.

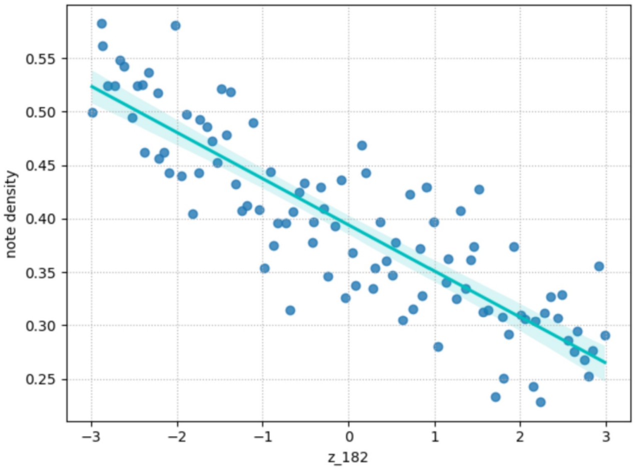


Figure 2: Linearity of the note density within the latent space

A further inspection of the structure of embedding space is performed using MDS, as described in Section 3. This low-dimensional representation allows immediate visualization of the relation between the structure of the latent space and the expected characteristics of the output. Fig. 3 shows the MDS plots associated with each of the four attributes. This representation is a powerful tool for inferring the output attributes starting from the sample position in the latent space: the color assigned to each region reflects the expected characteristics of the samples in that region, measured according to the different definitions. No continuous gradients emerges from the MDS, mirroring the weak navigability of latent space. Finally, Fig. 4 shows the distributions described in Section 3. The best results are associated with note density: each distribution, representative of a note density range, is coherently organized around the expected mean value. For the sake of visualization, we also plot the results associated with Toussaint which, on the other hand, doesn't satisfy the same requirements.

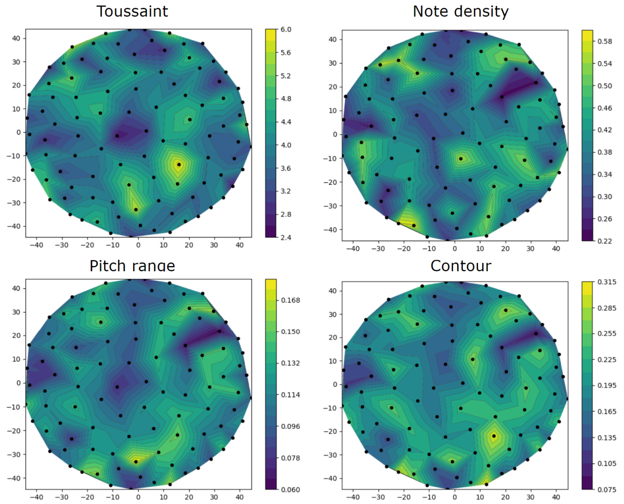


Figure 3: MDS of the latent space for different attributes

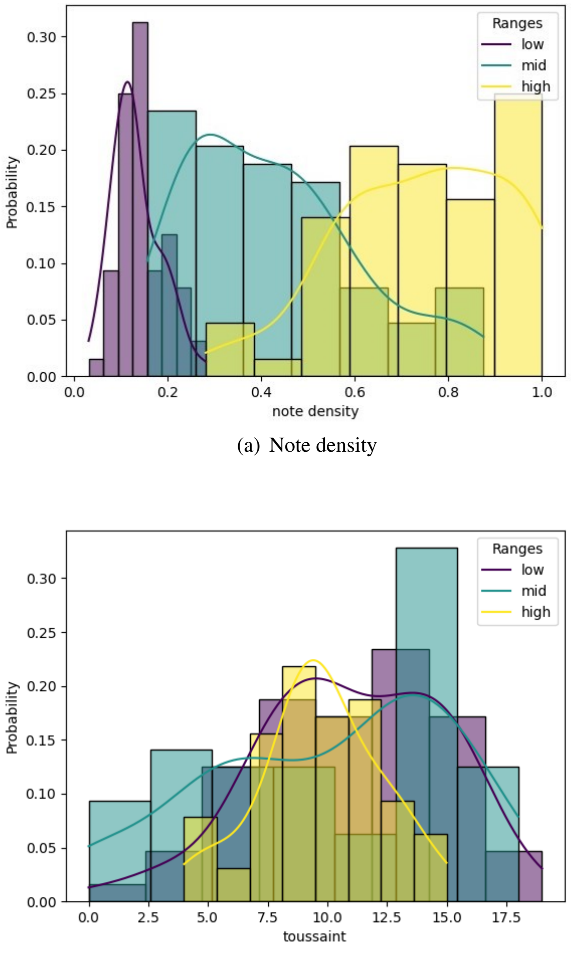
To proof any significant differences between the mean values of the measured high-level attributes, we perform a pairwise comparison of the distributions employing the Student's t-test [14]; the results related to the note density, reported in Table 1, show p-values $< 5\%$ thus confirming the presence of statistically significant discrepancies between the different distributions.

5. CONCLUSIONS

Our study explores the topological structure of the latent space learnt by the 2-bar pre-trained MusicVAE model by Magenta. We specifically seek the presence of any correlations between the characteristics of the embedding and a set of human-interpretable musical attributes.

We implement an efficient and organized mapping of the latent space, which allows a cross-check between the characteristics of the output and the allocation of the samples within the embedding. To achieve this we use LHS, a sampling technique suitable for semantically structure the latent space. The high-level attributes extracted from the model output are obtained analyzing its rhythmic complexity, note density, pitch range, and contour.

The latent space structure analysis identifies only one significant linear correlation between a dimension of the embedding and one of the considered attributes of the output: an inverse proportionality between the note density and the dimension 182 of the latent space. This result opens the door to possible human-AI interactions, such as the user's control over the number of notes of the generated musical sequence. Due to the depicted correlation between note density



(a) Note density

(b) Toussaint

Figure 4: Local attributes distributions

and contour, this would also influence the spread of the notes along the considered pitch domain.

The scarcity of linear relations between the latent space topology and the musical attributes of the output is corroborated by the results of further analysis, performed using MDS.

The results of our analysis show that the latent space lacks in navigability in relation to all the considered high-level attributes. Still, the characterization of the hyperspace in terms of musical attributes is possible and allows to head the decoding process toward definite target samples. To prove this, we identify three different local regions of the latent space associated with different attribute behaviours, we sample inside them, obtaining three sets of output attributes distributions. The Student's t-test on the distributions confirms that the distribution are certainly different if we consider the note density attribute.

Our study highlights a non-systematic navigability of the latent space: it is possible for deep generative models to spontaneously organize their latent space according to human-interpretable musical attributes, but this does not happen in a predictable way. To overcome this lack and properly condition the decoding phase, other approaches are needed such as the implementation of conditional VAEs or post-hoc techniques [3].

Attributes	Pair		
	low/mid	low/high	mid/high
toussaint	$2.214 \cdot 10^{-1}$	$2.768 \cdot 10^{-2}$	$6.610 \cdot 10^{-1}$
note density	$1.429 \cdot 10^{-19}$	$1.060 \cdot 10^{-38}$	$1.029 \cdot 10^{-17}$
pitch range	$5.314 \cdot 10^{-3}$	$1.698 \cdot 10^{-1}$	$1.757 \cdot 10^{-1}$
contour	$1.717 \cdot 10^{-1}$	$1.203 \cdot 10^{-9}$	$1.045 \cdot 10^{-1}$

Table 1: Student's t-test

6. REFERENCES

[1] D. P. Kingma and M. Welling, "Stochastic gradient vb and the variational auto-encoder," in *Second international conference on learning representations, ICLR*, vol. 19, 2014, p. 121.

[2] H. Dang, L. Mecke, F. Lehmann, S. Goller, and D. Buschek, "How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models," 2022.

[3] J. Engel, M. Hoffman, and A. Roberts, "Latent constraints: Learning to generate conditionally from unconditional generative models," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Sy8XvGb0-

[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022.

[5] S. Odaibo, "Tutorial: Deriving the standard variational autoencoder (vae) loss function," 2019.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[7] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, pp. 2673 – 2681, 12 1997.

[8] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," 2019.

[9] M. Stein, "Large sample properties of simulations using latin hypercube sampling," *Technometrics*, vol. 29, no. 2, pp. 143–151, 1987. [Online]. Available: http://www.jstor.org/stable/1269769

[10] R. Bergmann, J. Ludbrook, and W. P. J. M. Spooren, "Different outcomes of the wilcoxon-mann-whitney test from different statistics packages," *The American Statistician*, vol. 54, no. 1, pp. 72–77, 2000. [Online]. Available: http://www.jstor.org/stable/2685616

[11] A. Pati and A. Lerch, "Attribute-based regularization of latent spaces for variational auto-encoders," 2020.

[12] A. Mezza, M. Zanoni, and A. Sarti, "A latent rhythm complexity model for attribute-controlled drum pattern generation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, 02 2023.

[13] G. T. Toussaint, "A mathematical analysis of african, brazilian, and cuban clave rhythms," in *A Mathematical Analysis of African, Brazilian, and Cuban Clave Rhythms*, 2002.

[14] W. Haynes, *Student's t-Test*. New York, NY: Springer New York, 2013, pp. 2023–2025. [Online]. Available: https://doi.org/10.1007/978-1-4419-9863-7_1184