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 artistic applications such as image, speech, text and music generation. Selective manipulation of data attributes is one crucial aspect of these generative processes and it usually consists of incorporating the conditioning within the model or of implementing post-hoc learning latent constraints. However, the possibility of refining the conditioning process still represents an active area of research. It is with this mind that we develop the proposed analysis. Our study is focused on investigating the presence of any correlations between the topological structure of the latent space and high level features of the output. We examine a 2-bar pretrained model of MusicVAE, a recurrent Variational AutoEncoder by Magenta. As for the above-mentioned correlation, it has been investigated in terms of four different human-interpretable musical attributes. Latin Hypercube Sampling allows the regularization of the latent space, which enables explicit control of the output's features respect to the locations of the samples within the embedding. Our analysis aims at supplying a valid starting point for the development of controlled navigation techniques of the latent space in deep generative models, thus enabling a way less expensive conditioning process.

Index Terms— Variational autoencoders, Latent space topological structure, High level features, Latin Hypercube Sampling

1. INTRODUCTION

Deep latent variable models have gained an increasingly important role in creative applications of machine learning [1] [2]. Among these, Variational Autoencoders (VAE) stand out: the continuity of their latent space enables interpolation, which provides a creative way to explore artistic ideas. Many researchers have recently focused on improving the ability to condition these models, operating both upstream and downstream the training phase. In this regards, a popular approach consists of developing user-specified constraints during training, either by training on a curated subset of data or with conditioning variables. Despite being suitable to disparate application fields, they're restricted to the presence of enough labeled data available and require expensive model retraining for each new set of constraints. Another important method involves post-hoc learning latent constraints and enable to condition generation without retraining the model, by applying behavioral constraints on a embedding space, considered as a source of prior knowledge [3]. The proposed work fits into this context, with an additional mindset that proves the validity of the above described methods for the conditioning of the embedding. In particular, it focuses on the structural analysis of the latent space of a trained generative model and aims to identify a topological structure within the latent space itself.
Several supervised and unsupervised methods for editing semantics have already been refined, but they either require dedicated training phases or ex post estimates, thus limiting their applicability to sectoral fields. Our project is intended to outline a possible simulacrum for detecting the presence of intrinsic correlations between high level features, thus inquiring the possibility of consciously navigating the embedding and consequently influence the decoder's output. For our analysis, we worked on the pre-trained models of MusicVAE by Magenta and focused on spotting such correlations by measuring four different musical attributes; however, the versatility of the proposed method makes it easily suitable for detecting different kinds of patterns across the latent space, both in terms of measured features and analysis techniques.

2. RELATED WORK

Our analysis concerns the latent space structure of an artificial neural network architecture.
The deep generative model we worked on consists of a pre-trained configuration of the β-VAE architecture of MusicVAE by Magenta [4].
MusicVAE is a recurrent Variational Autencoder (VAE): a latent space model, capable of learning the fundamental characteristics of a training dataset, i.e. to translate the variation of real data in a lower-dimensional space. The main feature enshrining the difference between VAEs and simple autoencoders is the nature of the latent code z, obtained by the dataset compression: the VAE learns codes not as single points, but as soft ellipsoidal regions in latent space (posterior distribution over z), forcing the codes to fill the space rather than memorizing the training data as isolated codes [5].
The main advantage of VAEs with respect to regular autoencoders is their intrinsic regularization of the latent space, feature of major importance in the generative process.
The core of a VAE is the combination between encoder and decoder: the former, given a datapoint $\mathbf{x}$, produces a approximate posterior distribution $q(z|x)$ (e.g. a Gaussian) over the possible values of the code $\mathbf{z}$ from which the datapoint $\mathbf{x}$ could have been generated; the latter, given a code $\mathbf{z}$ produces a distribution $p(x|z)$ over the possible corresponding values of $\mathbf{x}$. Being θ and ϕ the weights which parametrize $q(z|x)$ and $p(x|z)$ on the neural networks, we'll refer to them as $q_\theta(z|x)$ and $p_\phi(x|z)$ [6] [7].
Since we're dealing with VAE, the deterministic bottleneck layer z is replaced with a stochastic sampling operation: instead of directly learn the latent variables, it learns a mean μ and a standard deviation σ which parametrize a probability distribution for each of

these latent variables. The probabilistic representation of the latent space z is therefore obtained as the result of a stochastic sampling operation starting from this distribution. 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 we adopt the reparametrization trick, which consists of recasting this statistical expression in a different way while preserving its meaning, thus allowing backpropagation. The sampled latent vector z is considered as the sum of a fixed μ and σ vectors, as in (1)

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

Thanks to this reparametrization, the stochastic sampling does not occur directly in the bottleneck layer z, allowing the VAE to be trained end to end [8].
The previously described high-level architecture has several implementations, according to the selected configurations. In our analysis we worked on a pre-trained 2-bar model, realized exploiting the combination of a Long Short Time Memory (LSTM) encoder with and a "flat" LSTM decoder. The idea at the basis of the functioning of the bidirectional encoder is to split the state neurons of a regular RNN in 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. As for the training, since there are no interactions between the two types of state neurons, the bidirectional neural network can be treated with the same algorithms as a regular unidirectional RNN and it can be unfolded into a general feedforward network [9]. The final state vectors are then concatenated and fed into two fullyconnected layers to produce the latent distribution parameters μ and σ. The two-layer bidirectional LSTM [10] [9] network state size is 2048 for all layers and produces an embedding having 512 latent dimensions. As regards the decoder, it is composed by a two layer RNN composed of 256 LSTM cells, producing a single categorical output [4].

3. METHODOLOGY

In this section, we submit an analysis technique of the latent structure of a trained generative model, enabling the definition of correlation patterns associated to high-level features. Initially we define a regular and replicable sampling framework within the latent space, which substitutes the canonical randomic sampling carried out by MusicVAE in inference mode. In particular, we implement a Latin Hypercube Sampling (LHS), for generating a near-random sample of parameter values from a multidimensional distribution. Despite this representation still includes a stochastic component, it turns out to be suitable for our model, due to its high dimensionality. The main advantage of LHS is that it simultaneously stratifies on all input dimensions to improve the coverage of the input space [11]. We also explored the possibility of a regular D-dimensional sampling grid, but we had to discard it due to the unacceptably high number of points: a total of D^N points with N points for dimension for a space of N dimensions. Fig. 1 shows the grid-like structure obtained with LHS, which provides an ordered and more efficient mapping of the plane with respect to the other sampling methods displayed. Each point of the Latin hypercube represents the centroid of a region in the hyperspace and corresponds to the mean of a Gaussian distribution; we proceed performing another sampling from this last distribution and use the collected samples as input for the model's decoder, which outputs a MIDI file in a deterministic way. Then, we make an high level feature analysis on the MIDI files, trying to identify a link between the ultimate results and the starting samples' coordinates on the Latin hypercube.

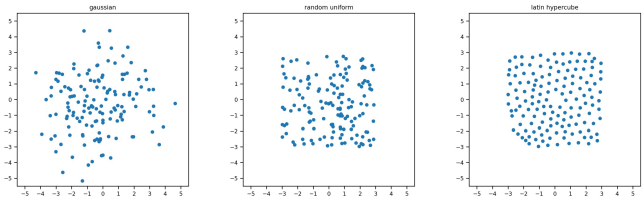

Figure 1: Comparison of different sampling methods.

Investigating the navigability of the latent space, i.e. the presence of any correlation between the location of the samples in the hypercube and the output's characteristics, represents the basis for a driven decoding process. To study the possible presence of link between human-interpretable musical attributes of the output and the structure of the latent space, we perform different statistical analysis. We start gauging the attributes of the generated MIDI file, then collect the results according to the outcomes. For each measured quality we compute the mean value of the samples associated to each point of the Latin hypercube, so as to obtain N averages for the N Gaussian distributions. Dealing with an high dimensional space makes the visualization of the results way less straightforward. To overcome this issue we implement a Multi-dimensional Scaling (MDS), leading to a bi dimensional representation of the data in which the distances respect well the distances in the original high-dimensional space [12]. We also conduct a second statistical investigation; we start identifying the highest, mid and lowest mean values out of the N ones and center a Gaussian distribution at each of these. Secondly we feed the samples derived from each distribution into the decoder and evaluate the attributes of the associated MIDI files: this involves the generation of three additional distributions, built according to the output's characteristics. Finally, we compare the ultimate retrieved distributions, displaying them on a histogram together with the relative Kernel density estimation. The presence of empty MIDI files doesn't affect the validity of the results and is simply discarded for the computation of the results.

4. EVALUATION

In this section, we discuss the details of our analysis for depicting any link between the latent space regularization and the output characteristics. First we place our intervention within the pipeline of MusicVAE. Secondly we present our choices for the regularization of the latent space. Then we list the definitions we adopted for the evaluation of high level features. Finally, we report and comment the obtained results.
To carry out our study, we extend musicVAE introducing the possibility to constrain and control the latent sampling, without modifying the model itself. The model is trained with the data from the MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization) dataset [13], composed of about 200 hours of paired audio and MIDI recordings of virtuous piano performances. We conduct our tests on a pre-trained checkpoint of MusicVAE, which generates melodies of 2 bars. The dimensionality of the regularized latent space depends on the selected configurations: for our analysis the model has a latent space of $D = 256$ dimensions.
For implementing the LHS described in Section 3, we set a total

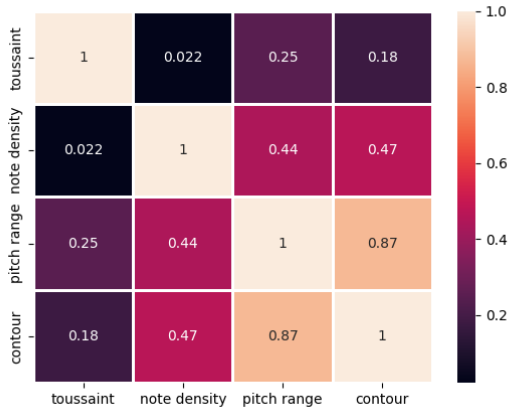

Figure 2: Pearson correlation between high feature values

of 100 points g_i, with $i = 1, \ldots, G$, collected in the vector $\boldsymbol{g}$ and representing the centroids of the regions of the hyperspace. For each g_i, we compute $X = 64$ multivariate Gaussian distribution $\mathcal{N}_D(\mu, \sigma^2)$, where $\mu = g_i$ and $\sigma^2 = 1$. By doing so, we obtain a total of $S = X \cdot G$ samples, whose coordinates are stored in the latent codes array $\boldsymbol{x} = \left[x_{1,1}, \ldots x_{G,X}\right]$. The latter, despite embracing the totality of samples, retains the ability to discern the region of origin of each one. The vector $\boldsymbol{x}$ is fed into MusicVAE's decoder that generates S MIDI files.

For investigating the navigability of the latent space, we analyze each MIDI output exploiting four different high level features. In prticular, we focus on the evaluation of rhythm and pitch, which are primary characteristics of a monophonic melody. To properly quantify these attributes, we adopted the definitions reported below [14].

- **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 to 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 [15].
- **Pitch range**: Corresponds to the normalized difference between the maximum and minimum MIDI pitch values, where the normalization factor R depends on the range of the dataset
- **Note density**: Counts the number of notes per measure normalized by the total length of the measure sequence
- **Contour**: Measures the degree to which the melody moves up or down and is measured by summing up the difference in pitch values of all the notes in the measure.

The results of each test are collected into four vectors of $\boldsymbol{a_s} = \left[a_1, \ldots a_S\right]$ elements.

4.1. Results and discussion

To properly interpret the values stored in the features vectors, we start depicting any correlation between the chosen attributes. To do so, we compute the Pearson correlation coefficient ρ between the assets of the collected attributes, reported in Fig.2. For the 2 bars model we get $\rho = 0.87$, between note density and contour. Such an high value suggests the presence of a strong correlation binding the two features; an higher number of notes tends to be spread on a wider pitch domain, so as to increase the contour.

Determining the presence of an intrinsic topological structure within the latent space corresponds to identifying a linear correlation associated to the output's characteristics along any direction of the latent space. To do so we refer to the vectors $\boldsymbol{g}$, $\boldsymbol{x}$, $\boldsymbol{a_s}$ defined in Sec.3 and proceed by taking into account one attribute at a time. We perform two different tests: we compute ρ between $\boldsymbol{x}$ and $\boldsymbol{a_s}$, along a defined direction of the hyperspace at a time; then, we repeat the same operation between $\boldsymbol{g}$ and $\bar{\boldsymbol{a}}_{\boldsymbol{s}}$ i.e. the vector containing the mean value of the measured features, associated to each region of the hyperspace centered in g_i. From the obtained values

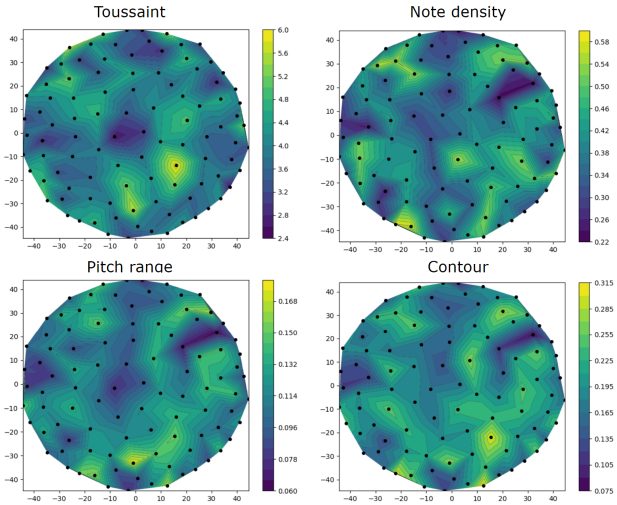

Figure 3: MDS of the latent space for different high level features

of ρ we can infer the presence of an intrinsic directionality within the latent space which influences the output characteristics in terms of note density. This trend is clearly displayed in Fig.4; the depicted a 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. A further inspection of the structure of the embedding is performed by using the MDS described in Section 3. This low dimensional representation allows an 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 to each of the four features. The color assigned to each region reflects the expected characteristics of the samples in that region, measured according to the different definitions: more dense colors correspond to higher values respectively positive (red) or negative (blue). Regions associated to similar values are much less spread for the note density and contour analysis; hence, we can strengthen an higher navigability of the latent space,i.e. the presence of a linear function orderly interpolating the attribute's values. This representation is a powerful tool for the inference of the output's attributes starting from the sample's position in the latent space. For example, it may lead to the implementation of a user-friendly model which catches the suitable sampling

coordinates within the latent space, according to the desired outputs' features. The spotted navigability of the embedding is not as immediate for the remaining high level features. To overcome this lack and properly conditioning the decoding phase, other post-hoc techniques are required [3]

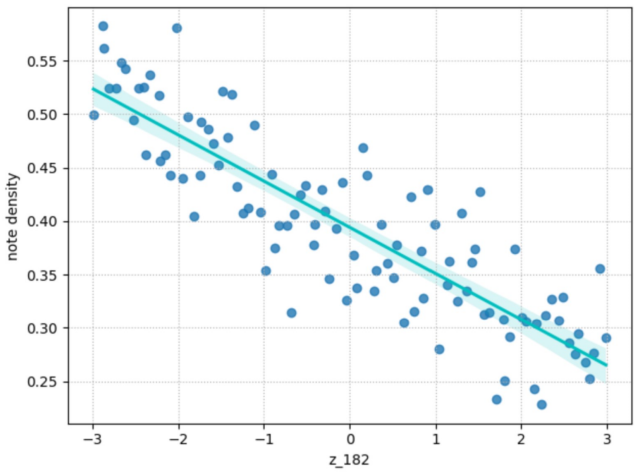

Figure 4: Linearity of the note density within the latent space

Finally we depict the unit-variance distributions described in Section 3. For the sake of visualization, we only plot the results associated to note density in Fig.5, which leads to the best results. Each distribution, representative of a note density range, differs from the others both in shape and peak's location. In support of this quali-

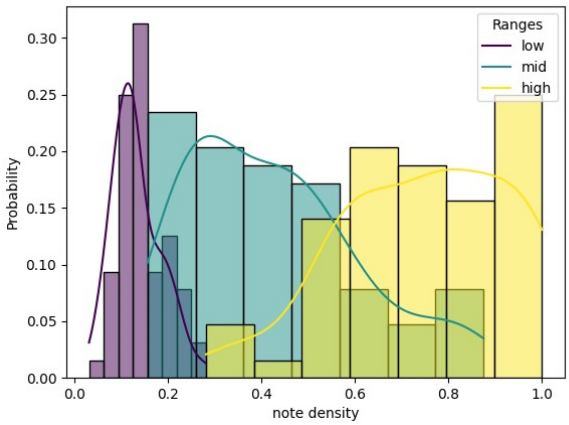

Figure 5: Note density histogram

tative analysis we perform two different tests, aimed to identify the presence of statistically significant differences between the mean values of the measured high level features. We compute the P-value in the Wilcoxon-Mann-Whitney test [16] and the Student's t-test [17], comparing the distributions in pairs; the results related to the note density, reported in Fig. 6, show P-values $< 5\%$ thus confirming the presence of statistically significant discrepancies between the different distributions. This also applies for the contour

	low/mid	low/high	mid/high
Toussaint	3.778E-01	2.735E-02	3.060E-01
Note density	4.009E-21	1.249E-22	4.684E-14
Pitch range	1.848E-02	5.783E-02	4.394E-01
Contour	1.978E-01	2.346E-15	4.241E-17

(a) Wilcoxon-Mann-Whitney test

	low/mid	low/high	mid/high
Toussaint	2.214E-01	2.768E-02	6.610E-01
Note density	1.429E-19	1.060E-38	1.029E-17
Pitch range	5.314E-03	1.698E-01	1.757E-01
Contour	1.717E-01	1.203E-09	1.045E-10

(b) Student's t-test

Figure 6: Statistical analysis of the results

measures (apart for the low/mid comparison), corroborating the correlation depicted in Fig.2.

5. CONCLUSIONS

In our project, we investigate the topological structure of the latent space of a deep generative model. We performed the analysis on MusicVAE by Magenta and focused on depicting any correlation between the characteristics of the embedding and several human-interpretable musical attributes. The proposed modus operandi relies on implementing a regularization of the latent space which allows a cross-check between the characteristics of the output and the sampling allocation within the embedding. To achieve this, we opted for LHS, a sampling technique suitable for semantically structure the latent space. As for the high level features, we analyzed the output in terms of rhythmic complexity, note density, pitch range and contour. By studying the latent space structure with respect to the output's note density, we identified a significant correlation enabling a feature-driven navigation of the latent space. The presence of a spontaneous topological structure is attested by the provided MDS plots. As regards the other investigated features, they didn't reveal any equally significant trend. However, the characterization of the hyperspace in terms of musical attributes allows a possible manipulation of the embedding, based on heading the decoding process towards definite target samples. Our work also concerns a statistical evaluation of the Gaussian distributions associated to the decoder's output. In particular, we identified three different ranges for each definitions and shaped the decoding process so as to obtain a probabilistic distribution for each rank. To acknowledge our qualitative analysis we also performed the Student's t-test and the Wilcoxon-Mann-Whitney test and verified the presence of statistically significant differences among the note density and contour distributions.

Our analysis, despite being conducted on an extremely detailed context, allows for great flexibility due to its intrinsic versatility. The very exact framework may serve as a starting point for investigating the presence of correlation between different perceptual definitions and the topological structure of the latent space, as well as for implementing non-linear navigation techniques of the latter, prompted by the output's desired attributes.

6. REFERENCES

[1] S. Carter and M. Nielsen, "Using artificial intelligence to augment human intelligence," *Distill*, 2017, https://distill.pub/2017/aia.

[2] D. Ha and D. Eck, "A neural representation of sketch drawings," *CoRR*, vol. abs/1704.03477, 2017. [Online]. Available: http://arxiv.org/abs/1704.03477

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

[5] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 10–21. [Online]. Available: https://aclanthology.org/K16-1002

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

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

[8] 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.

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

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

[11] ——, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[12] 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

[13] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," *CoRR*, vol. abs/1810.12247, 2018. [Online]. Available: http://arxiv.org/abs/1810.12247

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

[15] G. T. Toussaint, "A mathematical analysis of african, brazilian, and cuban clave rhythms," 2002.

[16] M. W. Fagerland and L. Sandvik, "The wilcoxon–mann–whitney test under scrutiny," *Statistics in Medicine*, vol. 28, no. 10, pp. 1487–1497, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3561

[17] 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