GPianoroll: a Deep Learning System with Human Feedback for Music Generation

Miguel Marcos, Lorenzo Mur, Rubén Martínez-Cantín

Afiliación: Grupo de Robótica, Visión por Computador e Inteligencia Artificial (RoPeRT)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: m.marcos@unizar.es

Abstract

We present a human-feedback-based evaluation pipeline for deep learning generative models, tested on music composition. Leveraging Bayesian Optimization and asking users to grade the generated songs, we efficiently find the point of a given model's latent space that yields each volunteer's favorite without retraining.

Introduction

Picture a music generation system that could adapt to your every whim, at any given moment. It might seem like we're not that far, but the truth is that music is a medium so heterogeneous that even the latest advances in technology struggle to cover it thoroughly. Deep learning models are usually expected to generalize, to adapt to unseen data. However, this usually comes at the cost of detail, so focusing on particular kinds of data without risking overfitting becomes hard. For cases where a user would want to draw one specific sample, like ours, this supposes a problem, especially if we can't have an objective and/or automatic evaluation of the generated samples. Each person has its own unique view on music, and profiling such abstract qualities is pretty costly for both developers and users.

Our pipeline

We propose a solution to adapt a single, general model to different users' musical tastes. We create a function mapping a model's input to a user's numerical evaluation, for which we find an optimum using Bayesian Optimization (BO). This technique eases music's costly evaluation due to several pieces not being possible to listen to at the same time. Also, for the first time in music-oriented BO works, we add the possibility of creating multi-instrument pieces playing several notes at the same time, which turns the problem into a high-dimensional scenario

Music composition model

We train a custom version of the gold standard music composition neural network MuseGAN [1]. The model is trained over the Lakh Pianoroll dataset which features rock music piano rolls, 3D matrices representing instrument, time and pitch. The model takes 128-dimensional Gaussian noise as input and outputs a ten-seconds-long song for Drums, Guitar and Bass. Training the model creates a latent space at the input where we will later draw samples that will result in new piano rolls.

Bayesian Optimization

We apply BO to our case by creating a target function using the Bayesopt library [2]. We take the input, forward pass it through the model, obtain the song and ask the user for a grade. These output grades build an approximation of the user's taste in the form of a Gaussian Process (GP). This way, we can find the global optimum efficiently by querying the GP where potential maxima are found. In order to accelerate the process, we perform dimensionality reduction on the function's input by adding Random Embeddings [3], reducing it to 10 dimensions. We guarantee that a multivariate Gaussian reaches the model's input by applying the Box-Muller transform to the uniform input of the BO algorithm, which also allows us to map free extra data points at the beginning, contributing to a better defined approximation with almost no overhead.

Human studies

Our pipeline's end is, ultimately, the user. We perform human-lead experiments where a supervisor is in charge of generating and playing the samples for volunteers to evaluate aloud. The first song played to users corresponds to an all-zero input, which lies in the highest-probability region of the latent space.

From there, we first perform exploration of the samples and then start the optimization algorithm. The whole process lasts for approximately half an hour, accounting for scheduled stops to rest the hearing and avoid biases.

Results

To evaluate our pipeline, we gathered a total of eight volunteers, analyzed the metrics for their favorite songs and asked them to answer a short Likert-scale survey, regarding the overall quality of the model, their selected favorite piece, and their own ability as judges.

Overall, the volunteers' answers were positive, indicating that the optimization system led to better compositions than letting the general model sample randomly, and that they standed by their given scores. Questions and answers are gathered in Table 1. The reported metrics show great variation, as depicted in Figure 1, and so do the samples, which can be listened to at our site[1]. Subject 4 is of special significance, as they managed to completely silence the guitar and bass tracks, an improvable case without human intervention.

Conclusion

We have implemented and tested a pipeline for human-feedback-based customization of deep learning generative models, applying Bayesian Optimization to the case of multi-instrument music compositions. The problem was challenging due to the high dimensionality of the problem and other issues native to the music domain. However, both the samples' metrics and the volunteers' opinions show that our method was successful, setting a stepping stone for future work in the field of Bayesian Optimization applied to music.

REFERENCES

[1]. DONG, H.W. and HSIAO, W.Y. and YANG, L.C. and YANG Y.H. MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. *In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. New Orleans, USA. 2018.

[2]. MARTÍNEZ-CANTÍN, R. BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits. *Journal of Machine Learning Research*, 2014, 15(11, pp. 3735-3739.

[3]. WANG, Z. and ZOGHI, M. and HUTTER, F. and MATHESON, D. and FREITAS, N. Bayesian optimization in a billion dimensions via random embeddings. *IJCAI International Joint Conference on Artificial Intelligence*, 2013.

Table 1. Aggregated answers (in %) for the Likert scale survey. Columns correspond to: Strongly Disagree, Disagree, No Preference, Agree and Strongly Agree.

Question	SD	D	NP	A	SA
Overall, I like the songs	0	37.5	50	12.5	0
I like my selected piece	0	0	0	75	25
I like my piece over the first	0	0	25	25	50
I'm OK with my scores	0	12.5	0	62.5	25

Table 2. Metrics for each volunteer's selected piece.

Metric	1	2	3	4	5	6	7	8
Tonal Distance	1.1	1	1.7	0	1.4	1.3	1.3	1.3
Pitch Classes (Bass)	2	1.8	1.3	0	1.8	2	1.5	2
Pitch Classes (Guitar)	4.8	5	3.3	0	4.8	5.3	3.8	7.3

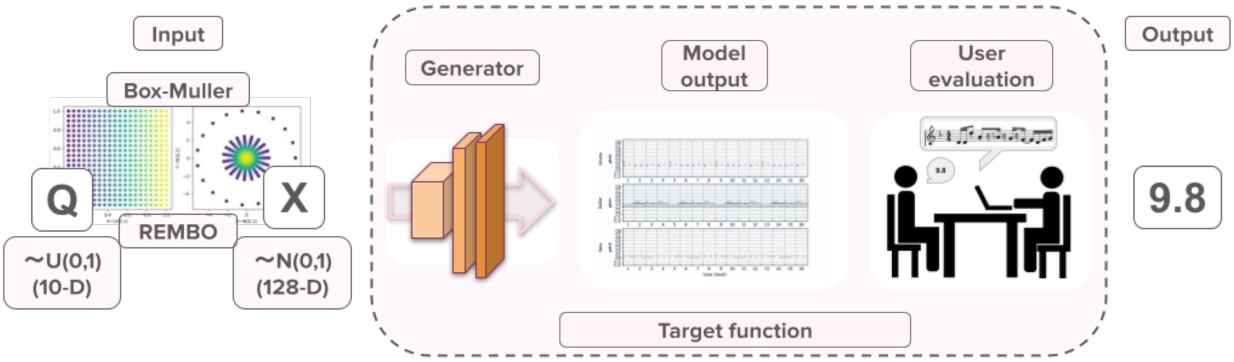

Figure 1. Overview of the whole optimization pipeline, from sampling, through generation and to grading.

[1] https://mikceroese.github.io/GPianoroll/