ICCC 2022 - Doctoral Consortium Research Summary

Berker Banar
Centre for Digital Music, School of EECS
Queen Mary University of London, UK
b.banar@qmul.ac.uk

Abstract

In this research summary document, first I describe and motivate my PhD project and then I explain recent work on four main pillars of my PhD, namely transformer-based symbolic music generation, music synthesis in the audio domain, explainable AI, and a quality-diversity-based evaluation strategy for generative music systems. In the last section, I mention some of our future work and how this doctoral consortium can be beneficial to my research.

Project Description

My PhD topic is 'Towards Composing Contemporary Classical Music using Generative Deep Learning', and is supervised by Simon Colton. This problem is important for a couple of reasons. Firstly, contemporary classical music is arguably the final frontier of human music culture as it provides unique abstractions as time series data. Advancing this research will enhance human creativity, especially in human-machine co-creative settings, pave the way for discovering new musical genres and enrich the evolution of our music culture in a wider spectrum. Also, composing in the style of contemporary classical music that has complex abstractions can be one of the creative benchmarks for artificial general intelligence studies, which aim to develop general-purpose systems that can learn a wide range of tasks concurrently, similar to human learning.

This project consists of various subfields such as symbolic music generation with transformer-based deep learning architectures, music synthesis in the audio domain using self-supervised representation learning for multi-modal settings, explainable AI techniques for controllable and transparent music generation models, and quality-diversity-based evaluation strategies for music generators encouraging out-of-distribution generation.

Recent Work

Transformer-based Symbolic Music Generation

In our work 'Generating Music with Extreme Passages using GPT-2' (Banar and Colton 2021b), we present a generative deep learning tool that can produce symbolic musical compositions with interesting and extreme passages using GPT-2 (Radford et al. 2019), and a novel method for controlling the generation of symbolic music. We utilise 4 different models of GPT-2 architecture. Each neural model is fine-tuned to a different loss value, and some of the models are not trained well on purpose, so that they generate passages that would likely not be composed by a person. Fine-tuned models are seeded with short musical excerpts to initiate the generation process. Using this to generate thousands of musical segments, we apply some musical analysis routines to categorise them in terms of long and fast melodic sequences, big interval jumps and atypical rhythmic figures, so that users can select segments in terms of how extreme they are using our user interface, then combine them into final pieces. In a second, fully automated, approach, we use the music at the end of one generated segment to seed the production of new segments, which can bring some level of continuity to pieces composed of chained segments in this way.

In this study, our main contributions are challenging the 'learning for perfection' idea, which might not necessarily be the best practice in creative settings, and composing pieces with particular arcs of extremity in terms of our musical metrics. We encourage the generation of passages of music outside the norm in terms of sequences of extremely fast, jumpy, or repetitive notes. In other contexts, these could be seen as defective and discarded as they don't reflect the musical distribution in the data particularly well. However, importing them into compositions in a controlled way could be musically interesting and a source of inspiration for new aesthetic territories. Also, as an interesting outcome, some of the generated pieces using our second iterative seeding approach are reminiscent of the music of Philip Glass and sequential music in general.

Also regarding symbolic music generation models and their evaluation specifically, in our papers 'Evaluation of GPT-2-based Symbolic Music Generation' (Banar and Colton 2021a) and 'A Systematic Evaluation of GPT-2-based Music Generation' (Banar and Colton 2022c), we systematically alter some typical training parameters of the GPT-2 model, namely the target loss level and training corpus, to investigate their effect on the generation process via controlled experiments. We then assess the music output by each model in terms of some musical metrics and statistical calculations. In these studies, we specifically focus on the evaluation aspect of music generators, as the evaluation of generative systems is important to provide insight

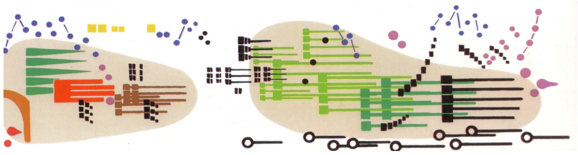

Figure 1: Graphic score fragment from Ligeti's Artikulation.

to generative music practitioners and improve the model architecture and training procedure. Moreover, if a generative music model is pre-trained with data from a different modality, e.g. natural language, evaluation in such cross-domain settings is crucial as cross-domain transfer learning brings its own challenges and there might be some limitations to transferred knowledge (Wu, Liu, and Potts 2021).

Our contribution in these studies has been to develop a methodology that consists of theoretically developing and implementing metrics of musical interest; a semi-autonomous approach to curating datasets of music, leading to more conclusive results from the experiments; varying the training of models; calculating and comparing summary statistics over the metrics; calculating and comparing KL-distance and overlapping area distances between training and generated distributions of musical samples, and using a web interface to visualise and interrogate the evaluation performed. We've learned from these studies that evaluation of generative systems in creative domains is important since typical loss functions as part of the training processes might not be reliable with peace of mind, and in transfer learning settings, evaluation is even more crucial. Also, as shown in our varying training level experiments, the well-trained model doesn't seem to learn musical chroma-related (tonality) features, which is an anomaly. Moreover, as shown in our varying training corpus experiments, depending on the characteristics of the training set, the model might not learn well specific musical attributes. In particular, average velocity, chroma and chords non-traditionality metrics are the ones to be more careful with in terms of poor learning performance. Furthermore, 20th Century music might be easier for the model to learn, or perhaps harder to confuse with the alternative classical music corpora in the study, which is further supported by our own aesthetic evaluation of the generated material using the Web App, where the generated material from 20th Century dataset sounds more musical, at least to our taste, subjectively. Using the generated material in these studies, we exhibit a piece, which is composed by a specialised GPT-2 model and arranged by ourselves, on our SoundCloud page [1].

Music Synthesis in the Audio Domain

Musical representations can be in various forms, such as audio, musical score, MIDI, or graphic score. Using an audio representation has the advantage of introducing expressive and textural elements in the sonic domain, where humans appreciate music. On the other hand, symbolic music representations such as musical score, MIDI and graphic score, are more practical for symbolic music generators, due to their confined form and being easier to manipulate. The music synthesis in the audio domain aspect of this PhD is implemented using self-supervised representation learning in multi-modal settings.

Figure 2: Edward Francis Burney, Amateurs of Tye-Wig Music (Musicians of the Old School), c.1820.

In our work 'Connecting Audio and Graphic Score Using Self-supervised Representation Learning - A Case Study with Gyorgy Ligeti's Artikulation' which is accepted to ICCC 2022 as a short paper (Banar and Colton 2022a), we present a novel self-supervised representation learning approach that can be applied to finding a mapping between audio and graphic scores in a generative context. Our approach consists of two variational autoencoder-based generators and a contrastive learning mechanism. We demonstrate this technique using György Ligeti's Artikulation, which is an electronic music composition with a graphic score that is depicted in Figure 1. In initial experiments, given manually designed graphic score excerpts in the style of Artikulation, we generated good quality audio correspondents with our model. Some ways of improving this system include data augmentation in both sonic and visual domains, and we suggest some use cases where an audio excerpt is illustrated in the style of the graphic score of Artikulation.

Explainable AI

Explainable AI increases the controllability of music generation models by making them more transparent and allows us to build trust with these systems. Even though explainable AI is often utilised in fields such as health care and justice, which are arguably about high-stakes decisions, it is still important to build trust with generative music systems as they shape the future of human culture. Also, explainable AI enables us to debug and diagnose machine learning models and improve their architecture.

In our paper 'Exploring XAI for the Arts: Explaining Latent Space in Generative Music' (Bryan-Kinns et al. 2021), we utilise a technique called latent space regularisation in a generative variational autoencoder model (Pati and Lerch 2019) (Pati and Lerch 2021) to force some specific dimensions of the latent space to map to meaningful musical attributes. Also, we provide a user interface featuring an in-

[1] soundcloud.com/user-330551093/untitled-by-gpt-2-and-anonymous-user-2022

teractive feedback loop via a web application to help people understand and predict the effect of changes to latent space dimensions on the generated music.

Quality-Diversity-based Evaluation Strategy

In our paper 'A Quality-Diversity-based Evaluation Strategy for Symbolic Music Generation' (Banar and Colton 2022b), we give a few musical examples of where loss-based and statistical methods fail and suggest some techniques for quality-based and diversity-based evaluation, jointly forming our evaluation strategy. We suggest methods such as using analytically computable metrics with statistical comparison, classifiers, metric learning approaches, the embedding space of the Wav2CLIP model (Wu et al. 2022), human audience evaluation, differentiable quality diversity strategy (Fontaine and Nikolaidis 2021) and a musical analogue of Self-BLEU metric (Zhu et al. 2018), which is typical in natural language processing applications. Moreover, we compare and contrast suggested methods in terms of some good practices and potential drawbacks and argue that quality-diversity-based evaluation approaches are highly appropriate for generative music models.

Future Work

In our future work, we are interested in further experimenting with transformer-based symbolic music generators to improve their performance, by exploring various data tokenisations, positional encodings, and attention mechanisms. Also, we plan to work on generative acousmatic music systems going beyond neural audio synthesisers, various explainable AI techniques such as concept whitening to perform surgery on generative music models, and quality-diversity-based analysis of our symbolic music and audio generators where out-of-distribution generation is encouraged to go beyond the limits of possibility space in a controlled way.

Being part of this doctoral consortium will provide a great opportunity to have feedback about various aspects of our work, such as evaluation of our systems, our experimental settings and new features that can be potentially added to these systems, and learn about new angles to the questions that we propose. Also, it will be insightful to guide our upcoming studies. Moreover, with the doctoral consortium, I hope to get mentoring and collaboration opportunities with the ICCC conference delegates, and to be inspired by the generative art and music projects that will be presented. As an inspiration for the future of this research, in Figure 2, we have a painting by Edward Francis Burney from 1820, in which a battle between modern and traditional music is depicted, where modern music is represented by Beethoven and Mozart, and traditional music is represented by Handel. As our musical culture has progressed a lot since those days, there will likely be new musical genres in the future, to which this research will contribute hopefully.

Acknowledgments

Berker Banar is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London.

References

Banar, B., and Colton, S. 2021a. Evaluation of GPT-2-based symbolic music generation. In *DMRN+ 16: Digital Music Research Network One-day Workshop 2021, Centre for Digital Music (C4DM)*.

Banar, B., and Colton, S. 2021b. Generating music with extreme passages using GPT-2. In *Evo* 2021*.

Banar, B., and Colton, S. 2022a. Connecting audio and graphic score using self-supervised representation learning - a case study with gyorgy ligeti's artikulation. In *International Conference on Computational Creativity (ICCC)*.

Banar, B., and Colton, S. 2022b. A quality-diversity-based evaluation strategy for symbolic music generation. In *International Conference on Learning Representations (ICLR) ML Evaluation Standards Workshop*.

Banar, B., and Colton, S. 2022c. A systematic evaluation of GPT-2-based music generation. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (EvoMUSART) (Part of EvoStar)*, 19–35. Springer.

Bryan-Kinns, N.; Banar, B.; Ford, C.; Reed, C. N.; Zhang, Y.; Colton, S.; and Armitage, J. 2021. Exploring xai for the arts: Explaining latent space in generative music. In *Conference on Neural Information Processing Systems (NeurIPS) eXplainable AI Approaches for Debugging and Diagnosis Workshop*.

Fontaine, M., and Nikolaidis, S. 2021. Differentiable quality diversity. *Advances in Neural Information Processing Systems* 34.

Pati, A., and Lerch, A. 2019. Latent space regularization for explicit control of musical attributes. In *International Conference on Machine Learning (ICML) Machine Learning for Music Discovery Workshop (ML4MD)*.

Pati, A., and Lerch, A. 2021. Attribute-based regularization of latent spaces for variational auto-encoders. *Neural Comput. Appl.* 33(9):4429–4444.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Wu, H.-H.; Seetharaman, P.; Kumar, K.; and Bello, J. P. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4563–4567. IEEE.

Wu, Z.; Liu, N. F.; and Potts, C. 2021. Identifying the limits of cross-domain knowledge transfer for pretrained models. *arXiv preprint arXiv:2104.08410*.

Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; and Yu, Y. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, 1097–1100.