A QUALITY-DIVERSITY-BASED EVALUATION STRATEGY FOR SYMBOLIC MUSIC GENERATION

Berker Banar & Simon Colton
School of EECS
Queen Mary University of London
London, E1 4NS, UK
{b.banar, s.colton}@qmul.ac.uk

ABSTRACT

Since human (audience) evaluation methods are challenging due to their logistic inconvenience, symbolic music generation systems are typically evaluated using loss-based measures or some statistical analyses with pre-defined musical metrics. Even though these loss-based and statistical methods could be informative to some extent, they often cannot guarantee any success for the generative model in terms of higher-level musical qualities, such as style / genre. Also, as another aspect of evaluation, diversity of the generated material is not considered for symbolic music generators. In this study, we argue that Quality-Diversity-based evaluation approach is more appropriate to value symbolic music generators. We give a few examples of where loss-based and statistical methods fail and suggest some techniques for quality-based and diversity-based evaluation, jointly forming a Quality-Diversity-based evaluation strategy.

1 INTRODUCTION

Music generation using machine learning is an active research field and one typical approach is to generate music in symbolic form, where fundamental musical attributes such as pitch, note start time, note duration and musical dynamics are represented numerically. Various learning architectures have been utilised in music generation studies including RNNs (Ji et al., 2020), VAEs (Roberts et al., 2018) and Transformers (Payne, 2019) (Huang et al., 2018). Evaluation of the performance of these systems is important for practitioners to choose a model that best fits their application and also to further improve these, and future, models along various axes. In most applications, generative music models are evaluated largely in terms of the final loss value at the end of training, with loss defined in terms of a statistical distance. In addition, there are some common post-hoc evaluation approaches, such as statistical comparison of training and generation sets, where being statistically closer is interpreted as having a better performing model (Yang & Lerch, 2020).

While comprehensive human (audience) evaluation studies are arguably the best approach for generative music systems, they can be difficult to conduct for researchers in terms of demanding logistics, especially for day-to-day evaluation of progress. Thus, researchers commonly use loss-based and statistical evaluation measures which are more convenient. However, these kinds of evaluation methods often lack higher level consideration of music, such as its style / genre. This is unsound in the context of generative music, where the generative output is multi-faceted, complex and fits into a rich cultural history. It is possible to train a model to low loss which produces either incoherent music in general or coherent music which doesn't reflect the genre of music it was trained on. Moreover, a model might be able to generate a set of musical excerpts that are statistically close to the training set, but this doesn't necessarily guarantee that the generated material stylistically reflects the set it was trained on.

As another drawback, existing loss-based and statistical evaluation methods do not take into account the diversity of the music generated by the models. In this paper, going beyond existing evaluation techniques, we argue that generative music systems need to be assessed in terms of both quality and diversity, hence once a model is trained, using Quality-Diversity search evaluation techniques (Pugh et al., 2016) is appropriate to estimate the value of the model. We suggest various techniques to

improve quality-based evaluation of generative music systems. Moreover, we propose some methods that can be applied to diversity analysis, jointly constituting a collection of methods for Quality-Diversity-based evaluation of music generators.

2 BACKGROUND

One common approach for quantitatively evaluating the performance of symbolic music generators is given in (Yang & Lerch, 2020), where the statistics of training and generation sets are compared using some pre-defined musical metrics. Here, musical metrics such as pitch count, pitch range, average pitch interval, note count and note length are calculated for each item in the training corpus. To be used for comparison purposes, a number of musical excerpts are generated from the model and the same musical metrics are calculated for each item in the generated set. Then, histograms for each metric are generated both for the training and generation sets, and distances between corresponding histograms are calculated using KL distance and an overlapping area method. If corresponding histograms of training and generated sets are closer to each other for model A than model B, then the authors claim that model A performs better than model B. While this approach is comprehensive and informative about the behaviour of models, it doesn't really provide an insightful picture of the generative system in terms of higher-level features such as style and genre.

Another measure of success for generative music models is having a low loss value, which is typical of almost any machine learning application. For generative tasks in creative domains, relying on the loss value might be problematic without any further analysis. To highlight this, we considered the successful and well-known symbolic music generation model called MuseNet (Payne, 2019), which can produce music in a given style. Using MuseNet, we generated two musical excepts in a jazz style for solo piano, without using an initiating seed. These examples can be found in Figure 1 in pianoroll format and audio files are on SoundCloud [1]. Based on our experience, even though these examples can be considered of good musical quality, they are not really in the style of jazz, due to the lack of syncopation, jazz phrasing and harmony. Instead, they are closer to classical music, stylistically.

In (Banar & Colton, 2022), we compare different training levels (in terms of loss value achieved) for a GPT-2 model in the context of symbolic music generation. We note that while a well-trained GPT-2 model generates music with chords, pitch count and average note duration similar to the musical training data presented, the music performs poorly when assessed with a chroma feature, which is a measure of tonality / atonality as defined in the work. In contrast, a poorly trained model performed much better in this respect.

These examples demonstrate that even though a model is successfully trained to a reasonably low loss value and seems to learn some musical features, it might not necessarily learn high-level musical qualities, such as style, or some musical features, such as chroma. This highlights the importance of quality-based evaluation. In addition, to the best of our knowledge, there is no symbolic music generation study, which focuses on the diversity of its generated material. Diversity is very important in generative tasks, because, in real-world cultural applications, the user normally requires a range of outputs to choose from.

3 A QUALITY-DIVERSITY-BASED EVALUATION STRATEGY

In this section, we first suggest some techniques that might improve a quality-based evaluation. These are presented in an increasing order of potential to capture high-level musical attributes, such as style and genre, but also in (subjectively) decreasing order of practicality. Secondly, we suggest using some of these techniques in diversity-based evaluation. The following approaches can be used for an assessment of the success of a generative music model, where we include the first and the last method to form a baseline in terms of practicality and capturing high-level musical attributes, respectively:

- **Analytically computable metrics with statistical comparison:** This approach consists of defining some analytical musical metrics and comparing the statistics of various sets using

[1] https://soundcloud.com/user-330551093/sets/examples-for-qd-evaluation

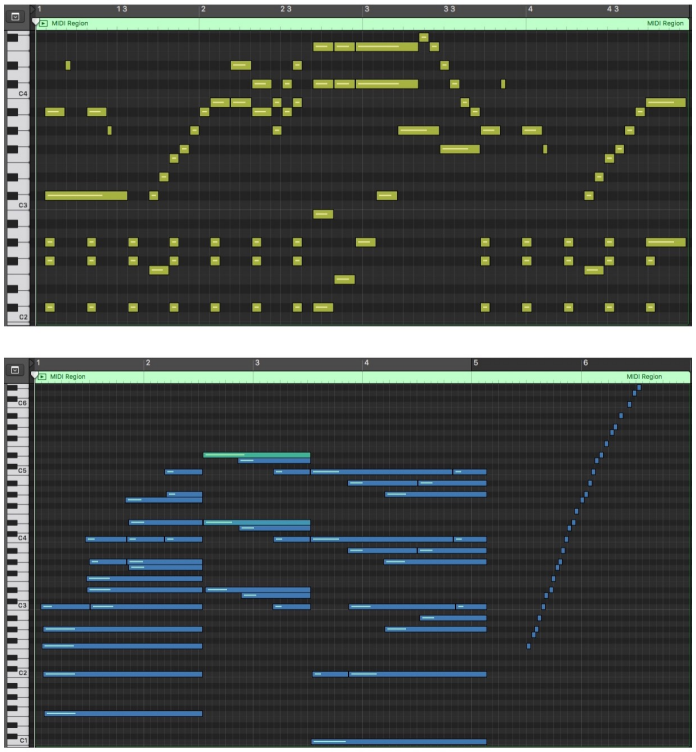

Figure 1: MIDI examples generated from MuseNet in the style of jazz. (Payne, 2019)

these metrics. One example would be the methodology explained in section 2, from (Yang & Lerch, 2020). In our list of techniques for quality evaluation, this approach would be the baseline in terms of computational cost, as computing analytical musical metrics and statistics is fairly easy compared to the techniques below. As a downside, arguably, this technique is not capable of capturing high-level musical qualities with analytically defined metrics.

- **Classifiers:** As we are interested in differentiating between various high-level musical attributes for quality evaluation, such as checking genre or style of a generated sample to see if it is aligned with the goal of the generator, a classifier-based approach might be useful for post-hoc quality evaluation. Since the embedding spaces of classifiers are not optimised in terms of the distances between different classes, these models might not necessarily perform well for distance / similarity related tasks (Lee et al., 2020). If there are too many different attributes or classes where an organised embedding space is needed, then a metric learning approach could be a better candidate.
- **Metric learning:** Metric learning approaches can be useful for quality-based evaluation methods, as they can provide a distance between different classes, in a scenario where we have a generated music sample and want to determine to which class it belongs, as well as its distance to other alternatives (Lee et al., 2020). Metric learning approaches may be suited to the given type of evaluation scenario, due to their organised representation space. One downside to metric learning is that the training process can take a long time, due to having duplet / triplet losses, where various sampling methods might be needed for efficiency, such as distance weighted sampling (Manmatha et al., 2017), or proxy-based training (Movshovitz-Attias et al., 2017).
- **Wav2CLIP:** Given the success of the CLIP's embedding space (Radford et al., 2021), Wav2CLIP (Wu et al., 2022) can be utilised to classify a musical sample and / or to get the distance between two different musical excerpts, which could be used for quality-based evaluation. Wav2CLIP maps audio recordings into the same embedding space as CLIP

does for images and text, and since it achieves competitive performance with various downstream classification tasks, this embedding space may be useful for quality-based evaluation of symbolic music generators. While this approach requires more computing resources, it might be quite informative about multi-criteria similarity. One important consideration, however, is that this approach works in the audio domain rather than symbolic music, which might require symbolic to audio conversion before using this approach, bringing its own challenges.

- **Human (audience) evaluation:** This approach is the most inconvenient one in terms of logistics, time and effort, but is arguably the best evaluation approach, as human perception is the ultimate goal and reference point. This technique would be the baseline for capturing high-level musical attributes.

Utilising these approaches depends on the computational resources available, and also the level of precision required for high-level musical attributes. Since there is a trade-off here, investigating the correlation between these techniques would be helpful, which might give an indication of the advantages / disadvantages of choosing one technique over another. One way of investigating the correlation between different techniques could be to use a correlation analysis matrix, which compares the performance of two different classifiers (A and B) using the items correctly classified by classifier A, items correctly classified by classifier B, items correctly classified by A but not B and items correctly classified by B but not A (Petrakos et al., 2000).

As we mentioned above, diversity is also important for the evaluation of generative music models, which should be balanced with consideration of quality. In other words, if the generation process is highly diverse in its output, but does not produce good quality results, then this case is not desirable. In an ideal scenario, if the generation process uniformly covers the characteristics of the training dataset, then hypothetically this would be the most diverse scenario given the training data. However, this might not be the case, and the model might be inclined towards a specific region in the possibility space. To overcome this problem, we need to consider the diversity of the learnt latent space by adding some extra criteria to the objective function as in (Fontaine & Nikolaidis, 2021). Here, they suggest a differentiable Quality-Diversity approach. In the context of music generation, this additional criteria can be about low-level musical qualities, such as pitch count or note duration; or about high-level musical metrics, such as style or genre. In the case of low-level qualities, analytical musical metrics might be enough. For high-level qualities, techniques such as classifiers, metric learning models or the Wav2CLIP embedding space can be utilised bringing their own computational complexities. Also, we can make an analogy between natural language generation and music generation domains and utilise a Self-BLEU-inspired metric for the diversity of music generators (Zhu et al., 2018). Self-BLEU is based on the BLEU (Papineni et al., 2002) metric, which evaluates the similarity between two sentences using n-gram precisions. Self-BLEU calculates a BLEU score for each sentence by considering it as hypothesis and the other sentences as reference in a piece of text, then averages these BLEU scores to obtain the Self-BLEU metric. In music generators, we can use musical fragments or themes instead of sentences and calculate a Self-BLEU score using these fragments, where BLEU metric is similarly based on n-gram precisions and uses musical attributes such as note number or note duration instead of words.

Furthermore, to go beyond the scope of the given training dataset and exceed the limits of its diversity, an out of distribution generation setting could be utilised (Möller et al., 2021), which might be achieved by adding small offsets to the fringy samples of the original distribution. This might enlarge the diversity limits of the model, which could create new opportunities with the potential to invent new genres / styles.

4 CONCLUSIONS

Symbolic music generators can be more appropriately evaluated using a Quality-Diversity-based strategy, which has not been used before in this context to the best of our knowledge. As shown in the examples, loss-based and statistical quality measures might not be sufficient to evaluate the performance of symbolic music generation models. To perform a Quality-Diversity-based evaluation, we suggest some techniques, which cover both quality and diversity aspects, and are presented in the scales of potential to capture high-level musical attributes and practicality. Also, we address some good practices and potential drawbacks regarding the proposed techniques.

In future work, we plan to empirically apply our Quality-Diversity strategy to evaluate and enhance performances of various symbolic music generation architectures. We are also interested in investigating the correlation between suggested techniques so that it will be possible to quantify the advantages of choosing one technique instead of another. Also, we would like to utilise this Quality-Diversity approach as a steering wheel for a symbolic music generation model, besides providing a post-hoc evaluation. Moreover, we want to expand on the diversity aspect of these networks so that it might be possible to invent new genres / styles, where inventing a new genre is arguably more creative than producing a piece within a genre through the lenses of computational creativity.

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.

We wish to thank the anonymous reviewers for their insightful comments.

REFERENCES

Berker Banar and Simon Colton. A systematic evaluation of gpt-2-based music generation. In *11th International Conference on Artificial Intelligence in Music, Sound, Art and Design (Evo-MUSART)*, 2022.

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

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer, 2018.

Shulei Ji, Jing Luo, and Xinyu Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions, 2020.

Jongpil Lee, Nicholas J. Bryan, Justin Salamon, Zeyu Jin, and Juhan Nam. Metric learning vs classification for disentangled music representation learning. In *The International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

R. Manmatha, Chaoxia Wu, Alex Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2859–2867, 2017.

Yair Movshovitz-Attias, Alexander Toshev, Thomas Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 360–368, 2017.

Felix Möller, Diego Botache, Denis Huseljic, Florian Heidecker, Maarten Bieshaar, and Bernhard Sick. Out-of-distribution detection and generation using soft brownian offset sampling and autoencoders, 05 2021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

C. Payne. Musenet. https://openai.com/blog/musenet/, 2019.

Michalis Petrakos, I. Kannelopoulos, Jon Benediktsson, and Martino Pesaresi. The effect of correlation on the accuracy of the combined classifier in decision level fusion. volume 6, pp. 2623 – 2625 vol.6, 02 2000. ISBN 0-7803-6359-0. doi: 10.1109/IGARSS.2000.859661.

Justin K. Pugh, Lisa B. Soros, and Kenneth O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3, 2016. ISSN 2296-9144. doi: 10.3389/frobt.2016.00040. URL https://www.frontiersin.org/article/10.3389/frobt.2016.00040.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. 03 2018.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip, 2022.

Li-Chia Yang and Alexander Lerch. On the evaluation of generative models in music. *Neural Computing and Applications*, 32, 05 2020. doi: 10.1007/s00521-018-3849-7.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, pp. 1097–1100, New York, NY, USA, 2018. doi: 10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.