View Reviews

Paper ID
1688

Paper Title
Generating Symbolic Music from Natural Language Prompts using an LLM-Enhanced Dataset

Track Name
ICASSP 2025 Main Tracks

Reviewer #1

Questions

2. Importance/Relevance
3. Of sufficient interest

5. Originality/Novelty
1. Has been done before; provides no new insight or understanding (implies reject, justify thoroughly)

6. Justification of Originality/Novelty Score (required)
The paper presents "MetaScore", a new dataset for symbolic music generation from natural language prompts. The dataset creation involves inferring missing tags by training a multi-tag classifier, and generating captions using large language models (LLMs). The approach of using LLMs for data generation is used extensively in multimodal understanding and generation tasks, including music domain (e.g. [1][2]).

On top of the dataset, a text-conditioned and a tag-conditioned music generation model are built. The task of generating symbolic music with free-form text input is not a fully researched topic. But the paper does not have specific novelty on the modeling side. The paper also does not contain the unique aspect of text to symbolic music v.s. audio music, which make the task itself looks like a simpler version of text-controlled audio music generation.

By the way, it seems that the same dataset is submitted as ISMIR late-breaking demo this year based on the info from this webpage (https://hermandong.com/publications/music_generation.html). It is questionable whether this should still be considered original.

[1] Doh, Seungheon, et al. "LP-MusicCaps: LLM-Based Pseudo Music Captioning." Ismir

2023 Hybrid Conference. 2023.
[2] Huang, Qingqing, et al. "Noise2music: Text-conditioned music generation with diffusion models." arXiv preprint arXiv:2302.03917 (2023).

7. Theoretical Development

2. Contains minor errors; provides barely sufficient insight or understanding

8. Justification of Theoretical Development Score (required if score is 1 or 2).

Since this is an empirical paper, there is not much theoretical development.

9. Experimental Validation

1. Insufficient validation (implies reject, justify thoroughly)

10. Justification of Experimental Validation Score (required if score is 1 or 2).

The experiment in this work has the following issues:

1. In the objective evaluation section, 3 models are compared against ground-truth on pitch class entropy, scale consistency and groove consistency. These metrics only imply whether the generated music is similar to the ground truth. There is no objective evaluation of whether the generated music conforms to the input tags or text. One possible way to evaluate would be to use the genre tagger to predict back tags and see if it align with the input tags.
2. Also for the objective evaluation, absolute values are used for the above metrics. In this way, it would be unclear if two values are on the opposite side of the ground truth value, which one is better. It would be better to report distance based metrics, such as KL divergence.
3. For subjective evaluation, the compared baseline (referred to as "BART-based") is trained on a different dataset. That dataset is in ABC-notation, and to the best of my knowledge, consists of mostly single track pieces. On the other hand, the proposed dataset seems to contain a substantial amount of multitrack pieces. Being multi-track would have an advantage in representing certain music styles. So it would be unfair to compare the baseline with models trained on the proposed dataset.
4. Just by listening to the audio demos, some of them do not seem to have strong relevance to the input tags or texts.

11. Clarity of Presentation

3. Clear enough

13. Reference to Prior Work

2. References missing

14. Justification of Reference to Prior Work Score (required if score is 1 or 2).

The creation of the proposed dataset is similar to that of audio domain music-text datasets but it lacks relevant citations (e.g. [1][2]).

[1] Doh, Seungheon, et al. "LP-MusicCaps: LLM-Based Pseudo Music Captioning." Ismir

2023 Hybrid Conference. 2023.
[2] Huang, Qingqing, et al. "Noise2music: Text-conditioned music generation with diffusion models." arXiv preprint arXiv:2302.03917 (2023).

15. Overall evaluation of this paper

2. Marginal reject

16. Justification of Overall evaluation of this paper (required)

This work proposes a new large scale symbolic music-text dataset and trained free-form text-to-music and tag-to-music generation models. It has the following major issues:

1. It seems that the dataset is published as 2024 ISMIR late-breaking demo, based on https://hermandong.com/publications/music_generation.html
2. Using LLM to generate pseudo text description is commonly seen in audio domain text-music datasets and other multimodal tasks (such as text-image). But this paper lacks related references
3. The objective evaluation does not prove the relevance of input texts/tags to generated music
4. The baseline model of subjective evaluation is not a fair comparison to the proposed models
5. Although the task of generating symbolic music with free-form text input is not a fully researched topic, the text-to-music and tag-to-music models in this paper are just transformer-based language models, which lack novelty.
6. The work also failed to emphasize the unique aspect of text to symbolic music v.s. audio music, such as enabling composers to do note-level editing. It looks like the task is just a simpler version of text-controlled audio music generation.

Reviewer #2

Questions

2. Importance/Relevance

3. Of sufficient interest

5. Originality/Novelty

3. Moderately original; provides limited new insights or understanding

6. Justification of Originality/Novelty Score (required)

Main new contributions:
1. This paper allows for free-form text-to-music generation in the symbolic domain, a task setup that has been heavily investigated in the audio domain, so this work is a step in the right direction and can fill an important gap in controllable symbolic music generation.
2. A new larger dataset is presented. The main advantage to this dataset is the availability

of detailed metadata. The main disadvantage is that only about 25% of the dataset is freely available/not subject to copyright restrictions. Seen in this light, the authors should reassess the advantages of this dataset wrt other symbolic music datasets already available in the public domain.

7. Theoretical Development

3. Probably correct; provides limited new insights or understanding

9. Experimental Validation

1. Insufficient validation (implies reject, justify thoroughly)

10. Justification of Experimental Validation Score (required if score is 1 or 2).

The experimental validation in this paper is extremely poor. While the methodology the authors use looks correct, the claims made in the results section do not follow from the metrics that have been reported, especially in Table 3. The authors seem to have compared different methods using only the mean while ignoring the information provided by the 95% CI.

Just to give one example, the authors claim "When contrasting MST-Tags-Small with MST-Tags, we observe that MST-Tags achieves better performance in coherence and arrangement, but we see a decrease in adherence..."

Doing a simple p-value analysis on the four metrics to compare MST-Tags-Small and MST-Tags will give p-values higher than 0.2, in some cases even as high as 0.5(!). This means that there is absolutely no significant difference and the conclusions that follow from these results do not hold.

Similarly, on coherence and arrangement, the difference between BART-based and MST-Text is not significant, whereas the authors write that " MST-Text outperforms the BART-based [10] approach, in terms of coherence, arrangement,..."

Such p-value estimates can provide a quantitative way to see what is already qualitatively clear from the numbers reported in Table 3. This is confusing because, for Table 1, the authors correctly note that the differences are not significant, but fail to do the same for Table 3.

Since the objective evaluation is quite limited, the paper could have perhaps used a better objective evaluation (Table 4) instead to support their claims, for example by using more suitable metrics from the music generation / text-to-music generation literature.

11. Clarity of Presentation

2. Difficult to read

12. Justification of Clarity of Presentation Score (required if score is 1 or 2).

I give a score of 2 because there are problems with the flow of the paper, which means that the paper would have to be reorganized / re-written significantly to make it readable. In many places, terminology/abbreviations are used that are not properly explained. This is particularly egregious when it happens in results tables, e.g. using “MST-Tags” or “MST-Text” in Table 1 but describing these terms only in Section 4, using “MetaScore-Genre” in

Table 2, discussing these results in Section 3B and 3C but introducing these terms only in Section 3D. Further, sometimes unexplained symbols are used, e.g. “*” in Table 1. More detailed table captions can resolve some of these issues.

13. Reference to Prior Work

3. References adequate

15. Overall evaluation of this paper

2. Marginal reject

16. Justification of Overall evaluation of this paper (required)

At a high-level, this paper is well-structured and the objective and methods are clear. However, upon closer examination, there are several issues with this paper which compound together to make it difficult to accept this paper in its current state. I detailed these problems above. In particular, in my view, the authors would have to do the following (which would change their current paper rather significantly):
(i) reorganize the paper to make it less difficult to read,
(ii) give a better objective evaluation to support their claims, and
(iii) fix the interpretation of their subjective evaluation to match what the numbers in Table 3 are showing.

Reviewer #3

Questions

2. Importance/Relevance

3. Of sufficient interest

5. Originality/Novelty

3. Moderately original; provides limited new insights or understanding

6. Justification of Originality/Novelty Score (required)

This paper presents a new large-scale user-anotated symbolic dataset

7. Theoretical Development

3. Probably correct; provides limited new insights or understanding

9. Experimental Validation

3. Limited but convincing

11. Clarity of Presentation

3. Clear enough

13. Reference to Prior Work

3. References adequate

15. Overall evaluation of this paper

3. Marginal accept

16. Justification of Overall evaluation of this paper (required)
In this paper, the authors presented a new dataset, which was collected from the public domain, with user-anotated metadata including genre, composer, user comment and pseudo captions.

The creation process consisted of several parts. To complete the missing part of metadata, the author proposed a new model, which was adapted from MMT, and trained this model on a subset with complete metadata to infer the missing part. Besides, the authors augmented the dataset by generating pseudo captions with large language models. Finally the authors verified the dataset by the user study, where 22 subjects were involved, to show the correctness in the statistical aspect.

Overall, the novelty of this paper is enough for proposing a new dataset. From my knowledge, there is no previous symbolic dataset that contains both large number of songs and paired metadata. The authors also train a model to predicts the missing values, which is a common approach in classical statistics but rarely mentioned in recent machine learning research.

However, I still have some doubts on the data quality. The author validated the dataset with a statistical method and showed the quality of inferred genre and generated captions are in the same level with the user-annotated data. Nevertheless, there is a lack of exploring the quality of the original data. As I know, MuseScore is open to anyone and has no censor for submissions. It would be better if this paper can add some comparison with other validated datasets to show the quality of crawled data.

In conclusion, I give a weak accept for this paper because of the uncertainty of the original data quality. I expect the validation I mention above can be added in the camera-ready version.