View Reviews

Paper ID
407

Paper Title
Multitrack Music Transformer

Track Name
Main Track: Audio and Acoustic Signal Processing

Reviewer #1

Questions

2. Importance/Relevance
3. Of sufficient interest

4. Novelty/Originality
3. Moderately original

6. Technical Correctness
3. Probably correct

8. Experimental Validation
4. Sufficient validation / theoretical paper

10. Clarity of Presentation
4. Very clear

12. Reference to Prior Work
3. References adequate

14. Overall evaluation of this paper
4. Definite accept

19. Detailed assessment of the paper (seen by the authors):
This paper contributes in 1) generation of multi-track symbolic music with improved inference speed, and 2) a quantitative measure of musical self-attention. Overall, this paper is very well written. The ideas presented in this paper are all clear and convincing. The insights given by this paper are reusable in future research in music generation.

Other comments:
1. Only 9 people participated the subjective test. Is the result reliable?
2. I guess that ideally the average generated sample length should be close to the average clip length in the training dataset. Can this be a evaluation measure?

Reviewer #2

Questions

2. Importance/Relevance
2. Of limited interest

3. Justification of Importance/Relevance Score (required if score is 1 or 2)

The topic - symbolic music generation using language models, feels a bit out of scope for a signal processing conference

4. Novelty/Originality
3. Moderately original

6. Technical Correctness
3. Probably correct

8. Experimental Validation
4. Sufficient validation / theoretical paper

10. Clarity of Presentation
4. Very clear

12. Reference to Prior Work
3. References adequate

14. Overall evaluation of this paper
4. Definite accept

19. Detailed assessment of the paper (seen by the authors):
This paper presents an approach for generating multitrack music from MIDI using a transformer model. While the topic of symbolic music generation with language models, is a bit of a mismatch for ICASSP, the paper is well-written with nice figures and examples. The proposed technique is computationally cheaper compared to previous work, but these savings come mainly from not having an autoregressive relationship between the different properties (e.g., pitch, duration) of a note. This savings does come at a cost, as the proposed method is outperformed by previous methods in both objective and subjective metrics. The authors also present an approach for analyzing the learned attention matrices, and find the model does learn to attend to regions that correspond to musically relevant pitch intervals and note durations.

-Section 3.1 - "our propsoed representation can represent 2.6 and 3.5 times longer music samples.." Since this is considered a major advantage of the proposed method, it would be better if this could be further broken down on what aspects of the representation enable this savings. For example, doesn't a majority of this gain come from reducing the number of MIDI instruments by half?

-Section 3.2 - What is meant by the topK sampling strategy?

The rebuttal generally answered many of the concerns I raised. I have upgraded my review accordingly.

Reviewer #3

Questions

2. Importance/Relevance
3. Of sufficient interest

4. Novelty/Originality
3. Moderately original

6. Technical Correctness
3. Probably correct

8. Experimental Validation
3. Limited but convincing

10. Clarity of Presentation

3. Clear enough

12. Reference to Prior Work

3. References adequate

14. Overall evaluation of this paper

2. Marginal reject

15. Justification of Overall evaluation of this paper (required if score is 1 or 2)

The more interesting part of this paper is the study of the musical self-attention. Unfortunately, this gets very little space in the end of the final section.

19. Detailed assessment of the paper (seen by the authors):

This paper is about a transformer-based approach to symbolic music generation. The authors propose a slightly different kind of encoding of the note seuqences as well as using a decoder-only transformer model with multi-dimensional input and ouput. Both objective scores and MOS-ratings show that the method performs slightly worse than state-of-the-art. Still, it is more efficient and can generate longer sequences.

View Meta-Reviews

Paper ID
407

Paper Title
Multitrack Music Transformer

Track Name
Main Track: Audio and Acoustic Signal Processing

META-REVIEWER #1

Not Submitted

META-REVIEWER #2

Not Submitted

META-REVIEWER #3

Not Submitted

META-REVIEWER #4

META-REVIEW QUESTIONS

1. Please provide your meta-review of the paper, incorporating the reviewer comments (seen by the authors).
Reviews commended the quality of writing for the paper, and the paper's novelty is also deemed sufficient. Please also note that topics related to symbolic music generation are very much in scope for ICASSP. On average, the paper is recommended for acceptance, noting that as suggested by reviewer #3 more focus could be put on the musical self-attention study.

META-REVIEWER #5

Not Submitted

META-REVIEWER #6

Not Submitted