View Reviews

Paper ID

5

Paper Title

Nested Music Transformer: Sequentially Decoding Compound Tokens in Symbolic Music and Audio Generation

Track Name

Papers

Reviewer #2

Questions

2. I am an expert on the topic of the paper.

Strongly agree

3. The title and abstract reflect the content of the paper.

Strongly agree

4. The paper discusses, cites and compares with all relevant related work

Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")

I'm missing a comparison (at least in the discussion of note-based encodings) with an approach similar to MuseNet, where each "compound token" is just a single atomic token in the vocabulary. I suppose the number of tokens would be too high if we wanted to encode all the features considered in this work, but this may be worth mentioning.

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.

Disagree

7. The paper adheres to ISMIR 2024 submission guidelines (uses the ISMIR 2024 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.

Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but

instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.

Strongly agree

9. Scholarly/scientific quality: The content is scientifically correct.

Agree

10. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen, otherwise write "n/a")

n/a

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.

Agree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.

Disagree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.

Disagree (Standard topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.

Agree

15. Please explain your assessment of reusable insights in the paper.

Reusable insights about the proposed method and related methods are limited due to a lack of clarity and detail. I'm also not sure how comparable NLL losses are across representations, so it may be hard to draw conclusions from that part of evaluation.

A clear insight, however, is that removing the conditional independence assumption of the compound token approach is beneficial for the quality of the generated music, and that this can be achieved at a cost that is lower than that of predicting each note feature as a separate "flat" token.

16. Write ONE line (in your own words) with the main take-home message from the paper.

A new architecture and token-based representation for more efficient symbolic music generation

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.

Strongly agree

20. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines

Weak accept

21. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.

The paper introduces Nested Music Transformer (NMT), an approach for generating symbolic music encoded as feature-rich token sequences (so-called "compound tokens", including note features such as pitch, metrical position, duration etc.) in an autoregressive way, feature by feature. This is a very much needed step from the original compound token approach, which made the assumption that each of the features of a note is generated independently. At the same time, this approach promises to be more computationally efficient than simply generating flat token sequences (i.e. representing each feature as a separate token).

NMT combines two Transformer decoders: a "main decoder", which operates on the level of notes (encoding the sequence of previously generated notes as a sequence of "hidden vectors") and a "sub-decoder", which, conditioned on this representation, predicts the features of a single note in an autoregressive way: an ordering is defined on the features and the sub-decoder predicts a feature given the values of the preceding features of the same note under this ordering. This idea is not new (papers with similar ideas are cited [4,13,8]). The main novelty claimed in this paper is to incorporate "cross-attention" in the sub-decoder.

The approach and the experiments seem well designed and executed. Two different variants of the proposed representation and four different architectures for the sub-decoder are explored and evaluated. In terms of NLL loss, NMT and its "lighter version" called CA outperform all other approaches except for a fully "unrolled" one (operating on flat token sequences – the least computationally efficient approach). A subjective evaluation is also performed with similar results, with the proposed approach being a clear winner (together with the flat approach). An experiment on audio tokens is also included, which is very interesting, even if it produced mixed results.

The main weakness of this paper is the clarity of the writing, especially in Section 3, which describes the approach. The notation is chaotic, descriptions somewhat unclear and incomplete, and figures not particularly easy to read. More specifically:
- The notation around attention is unclear. When the authors write "Cross-Attention(Query, K/V)", they should specify:
(1) What kind of attention? (multi-head scaled dot product?)
(2) If not multi-head, does it operate on the raw inputs, or are there linear projections for the values, keys, and queries?
(3) Is K/V_i a single vector, or is it a key/value pair (K_i, V_i)? I would suggest avoiding this notation if this refers to the input of attention before key/value projections.
- Writing "Cross-Attention(...)" and "Enrich-Attention(...)" makes it seem as if they're somehow fundamentally different modules, while in fact they are both instances of (cross-)attention, just with different inputs and weights (if they actually have weights). I would suggest just writing "Attention(...)" in both cases, with a subscript to denote weights. I would also suggest writing keys and values as separate arguments as is common practice.
- A more systematic notation should be chosen for tokens (discrete symbols) vs. their embeddings. Also, I would suggest to avoid using the term "features" for components of the compound token representation, as it may evoke a continuous feature. "Sub-tokens" would be a better name in my opinion.
- The description of the principle of the proposed cross-attention (CA) and feature enricher (part of NMT) should be made clearer. If I am reading this correctly, then, in my own words:
- In CA, the current output of the main decoder is combined with a positional encoding of current sub-decoder step (= feature index), then used as a query to attend across the embeddings of the previously generated sub-tokens ("features") to generate the next sub-token.
- In NMT, the embedding of the last generated sub-token ("feature") serves as a query to attend across the last w main decoder outputs – this is called "enrich-attention". Its output then serves as keys and values to "cross-attention" where the query is the same as in plain CA, i.e. the current output of the main decoder with a positional encoding.
I was not able to understand this from the text or from Figure 2 (difficult to follow), and it took considerable effort to get it from the equations, which are not completely precise either.

- I find the terms "enrich-attention" and "cross-attention" weirdly chosen. In my view, what is called "enrich-attention" is actually closer to a conventional cross-attention, where a decoder's hidden features attend to the outputs of some other module. What is called "cross-attention" here seems like the more exotic type of attention since it has a "static" query.
- I'm missing some motivation for this rather convoluted approach, as opposed to something more straightforward like [8].

Related to the last point, I find that the self-attention approach proposed in [8], if I understand it correctly, is actually more general than both the CA and SA proposed here: [8] also autoregressively predicts "sub-tokens", but seems to add the "hidden vector" to each sub-token embedding, so all available information influences keys, values *and* queries. In contrast, the proposed CA is more restricted (some information flows to keys/values, other information to queries). Moreover the self-attention strategy (SA) studied here is not equivalent to [8] because it instead seems to feed the hidden vector as a separate token, instead of adding it to all tokens. This means the present work does not properly compare to [8] and at the same time proposes an approach that is possibly less powerful and more convoluted, reducing the novelty and appeal of the approach. (However, I am not deeply familiar with [8], so I might be misunderstanding something.)

Regarding NMT, the main proposed model, I do not think it can be concluded from the results that it is better than the CA approach; in fact, it seems more or less on par in Table 1, and is consistently worse in Table 2. For this reason, it is not clear why NMT was chosen over CA for the subjective test and as the main model given that it's more complicated, slower and uses more memory.

More comments:
- Architecture hyperparameters like model size, number of layers, attention heads, the window size w, etc. are completely missing from the paper; only the total number of parameters is given.
- Section 2.1 is difficult to understand; unclear what "type" or "conti" means. On L118, it is unclear why the repetitiveness of beat position is a bad thing.
- In Table 1, you may consider highlighting the second-best, or the best non-REMI model for better readability.
- I'm not sure how useful Table 1 actually is. It is not clear how comparable NLL loss can be for different representations, especially with different vocabulary sizes.
- It is mentioned in the conclusion that the model "efficiently manages GPU resources and training processes". This claim seems a bit unsubstantiated, given that NMT is not exactly straightforward, and its efficiency is not elaborated upon in the paper (apart from numbers in Table 1, which do not make it look exceptionally efficient). In any case some analysis of

the computational efficiency would improve the paper.
- Multiple equations seem to be missing the index c.
- A more suitable name than "Catvec" would in my opinion be MLP or Feed-forward.

Reviewer #4

Questions

2. I am an expert on the topic of the paper.

Strongly agree

3. The title and abstract reflect the content of the paper.

Agree

4. The paper discusses, cites and compares with all relevant related work

Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")

I like the references and categorization. The authors may consider including Compose & Embellish [1] as a related work on token sequence ordering. C&E factorized symbolic-domain generation into melody and accompaniment/expression stages, which is conceptually similar to the "coarse first" setting of MusicGen.

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.

Agree

7. The paper adheres to ISMIR 2024 submission guidelines (uses the ISMIR 2024 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.

Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.

Strongly agree

9. Scholarly/scientific quality: The content is scientifically correct.

Agree

10. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen, otherwise write "n/a")

It could be my misunderstanding or writing issues, but I'll flag it here.

The "query" in sub-decoder cross-attention (leftmost part in Fig. 3) actually feels more like keys/values, as it seems to be "attended to" by f_c's, which produce the token outputs. Moreover, I don't quite understand why the query needs to be duplicated given that duplication does not provide additional information.

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.

Agree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.

Disagree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.

Disagree (Standard topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.

Agree

15. Please explain your assessment of reusable insights in the paper.

From "type 1st" and "pitch 1st" arrangements, we can know that even when the probabilistic factorization is exactly the same, which tokens are placed first in the sub-decoder can affect the attention/specialization of the model.

16. Write ONE line (in your own words) with the main take-home message from the paper.
Nested Music Transformer finds a good middle ground between fully parallel prediction of compound tokens and fully flattened token sequence, using a sub-decoder that sequentially predicts the elements in a compound token.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Agree

20. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines
Weak accept

21. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
Strengths
(S1) I like the comprehensive review and categorization of the existing methods aiming to strike a balance between efficiency and effectiveness (PianoTree VAE, MMT, CP Transformer).

(S2) From the NLLs reported in Table 1, the proposed model had solid advantages over CP Transformer and MMT-style fully parallel prediction. Ablations on cross-attention and feature enricher were also done to show their necessity.

Weaknesses
(W1) The writing/presentation issues in the manuscript are somewhat significant to an extent that impacts the understanding of the content. Aside from the confusion mentioned above in "Scholarly/scientific quality" part, I noticed two specific issues:
- The equations in Sec. 3.2. (cross-attention) and Sec. 3.2.1. (feature enricher) should be independent from each other as they are seperate modules as drawn in Figure 3. However, in the current manuscript, they are entangled (see Eqns (4-5) in Sec. 3.2 vs. Eqns (3-5) in Sec. 3.2.1).
- The monolithic Table 1 is hard to parse, and hence takes a lot of time and searching to realize that the proposed model is consistently stronger. For better clarity, I suggest breaking it down into 3 parts that echoes the organization of Figures 2 and 3 -- (i) tokenization schemes (REMI vs. CP vs. proposed Type 1st vs. proposed Pitch 1st), (ii) factorization (MMT-like parallel vs. CP-like partially-sequential vs. proposed fully-sequential), and (iii) sub-decoder design (self-attention vs. cross-attention vs. cross-attn + feature enricher).

(W2) The objective evaluation was rather shallow as only the NLL loss is examined. From the examples on the demo site, I think the proposed NMT model has a clear strength on multitrack music -- its generations are much better-structured and more musically pleasant due to clear sections and texture/instrumentation variations. More fine-grained metrics (e.g., variance in measure-level polyphonicity [1] or instrument counts) could be reported to manifest NMT's unique strengths.

Minor comments
(MC1) Typo in line 49: "tpye-note" --> "type-note"
(MC2) Line 183: ` h^{f_c}_{l} ` does not appear anywhere in the following paragraphs. Consider removing this.
(MC3) The equation numbering restarts from (1) multiple times. Please fix this.
(MC4) Typo in Eqn. (6) above line 219: Feature^{f-1} --> Feature^{f_{c-1}}

[1] Wu, Shih-Lun, and Yi-Hsuan Yang. "MuseMorphose: Full-song and fine-grained piano music style transfer with one transformer VAE." TASLP 2023.

View Meta-Reviews

Paper ID

5

Paper Title

Nested Music Transformer: Sequentially Decoding Compound Tokens in Symbolic Music and Audio Generation

Track Name

Papers

META-REVIEWER #1

META-REVIEW QUESTIONS

2. I am an expert on the topic of the paper.

Agree

3. The title and abstract reflect the content of the paper.

Agree

4. The paper discusses, cites and compares with all relevant related work.

Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")

n/a

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.

Disagree

7. The paper adheres to ISMIR 2024 submission guidelines (uses the ISMIR 2024 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.

Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed

method can also be applied to non-music domains if it is shown to be useful in music domains.

Strongly agree

9. Scholarly/scientific quality: The content is scientifically correct.

Agree

10. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chose, otherwise write "n/a")

n/a

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.

Strongly agree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.

Agree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated “Strongly Agree” and “Agree” can be highlighted, but please do not penalize papers rated “Disagree” or “Strongly Disagree”. Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.

Disagree (Standard topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.

Disagree

15. Please explain your assessment of reusable insights in the paper.

See below

16. Write ONE line (in your own words) with the main take-home message from the paper.

Symbolic and discrete token generation can benefit significantly from the right input representation and specifically designed final-stage decoders.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Agree

20. Overall evaluation (to be completed before the discussion phase): Please first evaluate before the discussion phase. Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines.
Weak accept

21. Main review and comments for the authors (to be completed before the discussion phase). Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
The authors present an LLM-based system primarily for generating symbolic music. The system employs a note-based input representation, where each note is represented by 8 concurrent features, which is in contrast to MIDI-like REMI encodings (i.e. one message per feature) and compound word encodings, which yield a different structure. After an embedding layer, the input representation is fed into a main generator whose output is given to a subdecoder. The proposed subdecoder uses a crossattention-based decoding mechanism (crossattention on latent vectors x output sequence - in contrast to regular RNN or self-attention decoders), which improves the results in experiments.

Overall, I am torn on the paper. On the one hand, a very solid technical paper bringing together a considerable number of technologies at scale while adding a number of novel ideas. My primary concern is that the paper employs a considerable number of non-trivial concepts, which for a lack of space cannot always be properly motivated and then have to be described at a rather superficial level - despite the author's best efforts to deal with that situation. This density makes the paper hard to read and thus a reader's value from it can be limited. It is appreciated that the authors plan to open source their contribution to alleviate some of these issues - yet for transferable insights a journal article providing more space might be the better format.

- Line 117: This marking seems to be missing?
- Line 127: This encoding would suggest that multiple concurrent notes can set different tempi. It does not become clear if the interpretation is similar to MIDI, where only the last note in the sequence defines the tempo, or whether this is done differently here. If the order in which concurrent notes are stored matters like this, the encoding could introduce an intrinsic uncertainty for how this is to be decoded into MIDI. As the model does not care about absolute time and thus any mismatch won't be regarded during training, this would not show up as a

problem during training.
- Line 145: It does not become clear why a shift like this leads to the model being more attentive to earlier features
- Line 159: "The token embedding summarizes the embeddings of each feature into a single vector." It does not become clear what 'the embeddings of each feature' are.
- Line 164-166: The lines contain multiple typos and should be rephrased
- Line 168: This section does not actually describe the main decoder
- Line 203: If I understand correctly, this process is repeated for each feature, i.e. one cross correlation. I would expect this to be much slower compared to the alternative approaches described in 3.3, yet Table 1 shows that the runtime per iteration for NB-PF + NMT is only ~80% slower compared to NB-PF + Catvec. Why is it not a lot slower?
- Line 211: The motivation for the enrich-attention process does not become clear, i.e. what problem is being solved?
- Line 319: "Instead of employing early stopping, dataset-specific dropout rates were applied to address overfitting concerns." While dropout is a regularizer and thus helps with overfitting, it does not become clear how it is a viable alternative to early stopping, especially since one still would need to check if overfitting occurs to set the dropout rates correctly
- Line 389: SOD should be spelled out.

22. Final recommendation (to be completed after the discussion phase) Please give a final recommendation after the discussion phase. In the final recommendation, please do not simply average the scores of the reviewers. Note that the number of recommendation options for reviewers is different from the number of options here. We encourage you to take a stand, and preferably avoid “weak accepts” or “weak rejects” if possible.

Weak accept

23. Meta-review and final comments for authors (to be completed after the discussion phase)

After the discussion, all reviewers agreed that on the one hand, this is an technically interesting paper that contributes to knowledge in the field. On the other hand, all reviewers also agreed that the writing quality is borderline. This leaves the papers very hard to read and could limit the value to the reader.

Overall, all reviewers agreed that a weak accept might be appropriate here.