

Literature Review

# Extending NLP Transformers for Structured Music Composition

---

By David Daniel, Danil Afonchikov, Ayhem Bouabid

# Article 1

Shih, Yi-Jen, et al. "Theme transformer: Symbolic music generation with theme-conditioned transformer." IEEE Transactions on Multimedia (2022).

## Problem Statement

- Conditional music generation without straying away from the conditioning signal like in prompt-based methods.

## Solution & Key findings

- Proposes the Theme Transformer. Themes are extracted using the clustering of melody-embedded segments of the original music. The experiment compares multiple variants of the Theme transformer, but all of them perform better than the prompt-based approach.

## Methodology critique

- The authors avoid rule based approaches in the Theme extraction, and instead train a separate model using contrastive learning to embed melodies for theme selection. This enforces the external validity of the experiment since this approach is more generalizable.
- Usage of Ablation studies, which help separate the effect of different introduced techniques
- Both objective and subjective metrics. We do not know the meaningfulness of the objective ones, or the agreement in the subjective ones.

# Article 1 Strengths and Weaknesses

Strengths	Weaknesses
<ul style="list-style-type: none"><li>- Usage of both objective and subjective evaluation</li><li>- Learned melody embeddings</li></ul>	<ul style="list-style-type: none"><li>- Only supports one Theme</li><li>- No generation from scratch</li><li>- Considers only melody for the Theme</li><li>- Fixed theme segment size</li></ul>



# Article 2

Hsiao, Wen-Yi, et al. "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 1. 2021.

## Problem Statement

- At the representation level, music is tokenized in a similar way to natural language but with different token categories, which are treated equally by the model.

## Solution & Key findings

- Special attention heads for each token type, and the tokens are grouped to create compound words of multiple tokens of different token categories. This reduces the sequence length and improves performance
- Faster training, ability to model longer sequences, better performance.

## Methodology critique

- Used both objective and subjective evaluation
- Only means are provided for the subjective evaluation which are not necessarily very representative
- The authors add a Graph interpretation for their model
- The evaluation includes model resource consumption

# Article 2 Strengths and Weaknesses

Strengths	Weaknesses
<ul style="list-style-type: none"><li>- Extensive evaluation with music objective and subjective metrics, and resource utilization metrics</li><li>- Theoretical formalization of the model</li></ul>	<ul style="list-style-type: none"><li>- Compared against a single baseline representation (REMI)</li></ul>



# Article 3

Wang, Ziyu, and Gus Xia. “MuseBERT: Pre-training Music Representation for Music Understanding and Controllable Generation” ISMIR. 2021.

## Problem Statement

- Representing polyphonic music as a sequence and embedding it is a challenge.

## Solution & Key findings

- Proposes MuseBERT, which uses BERT and a generalized relative position encoding to implement Attention for polyphonic music. The findings show results closer to human-composed music than the baseline.

## Methodology critique

- Mathematical foundation for the model effectiveness
- Multiple baseline comparisons and ablation
- Objective and subjective evaluation

# Article 3 Strengths and Weaknesses

Strengths	Weaknesses
<ul style="list-style-type: none"><li>- Usage of both objective and subjective evaluation</li><li>- Learned melody embeddings</li><li>- Subjective scores with the training data as a baseline</li></ul>	<ul style="list-style-type: none"><li>- Uses generic metrics for subjective evaluation (Musicality, Naturalness ..)</li><li>- Very low-level music objective metrics (average note duration)</li></ul>



# Article 4

Sulun, Serkan, Matthew EP Davies, and Paula Viana. "Symbolic music generation conditioned on continuous-valued emotions." IEEE Access 10 (2022): 44617-44626.

## Problem Statement

- Conditional models use pre-defined musical features, but music is about the emotion that it delivers, not the exact features that it has. One of the prominent models of human emotion is the arousal-valence model. The goal is to use this model to generate music.

## Solution & Key findings

- The paper starts by collecting a dataset that pairs music in the MIDI format to continuous-valued arousal values using the Spotify Dev API. Then the authors pre-train a Music Transformer model to unconditionally generate music, then fine-tune it on the collected dataset. The authors compare three different approaches to incorporating the conditioning signal: Either through discrete tokens, continuous-added, or continuous-concatenated tokens. The continuous-concatenated approach yields the best results.

## Methodology critique

- The evaluation uses only an objective evaluation that is not domain-specific. The model is evaluated through NLL and accuracy, and then through the performance of a model that predicts the arousal-valence values from the final output.
- There is no baseline model that uses other types of conditional generation.



# Article 4 Strengths and Weaknesses

Strengths	Weaknesses
<ul style="list-style-type: none"><li>- Data collection, the study offers a new dataset to the community</li><li>- Pre-training of a generic music generator to re-use weights</li></ul>	<ul style="list-style-type: none"><li>- No subjective evaluation</li><li>- No baseline</li><li>- Lacking domain-specific objective metrics</li></ul>



# Article 5

Qin, Yang, et al. "Bar transformer: a hierarchical model for learning long-term structure and generating impressive pop music." *Applied Intelligence* 53.9 (2023): 10130-10148.

## Problem Statement

- Lack of overall structure in the music generated by Transformers, and autoregressive models in general because of the iterative nature of decoders. The goal is to find a hierarchical generation method to enforce structure in the generated music.

## Solution & Key findings

- Proposes the Bar transformer, which has a hierarchical encoder and a decoder that cross-attends to multiple encoding levels at each layer. It performs better in terms of repetition and KL-divergence w.r.t training data.

## Methodology critique

- Several assumptions are either fully justified or not well formulated
- The evaluation is based on small datasets and hence hinders the generality of the findings

# Article 5 Strengths and Weaknesses

Strengths	Weaknesses
<ul style="list-style-type: none"><li>- Comparison to multiple baseline models</li><li>- Going beyond the standard text generation metrics</li></ul>	<ul style="list-style-type: none"><li>- Why not try different numbers of levels for the hierarchical generation?</li><li>- The authors' approach might fall under the umbrella of “feature engineering”</li><li>- The objective evaluation relies on “Repetition”, which is assumed to be a reliable proxy for structure</li></ul>



# Article 6

Qin, Yang, et al. "Score Images as a Modality: Enhancing Symbolic Music Understanding through Large-Scale Multimodal Pre-Training." *Sensors* 2024, 24, 5017.

## Problem Statement

- Music might be too complex of a problem to be tackled with a single Mode model.
- The need for a more holistic approach that integrates visual and symbolic representations of music to capture its structural complexity.

## Solution & Key findings

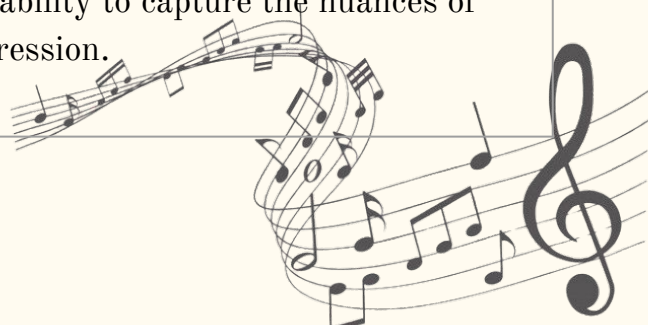
- The authors propose the Score Images as a Modality (SIM) model, which integrates music score images alongside MIDI data in multimodal pre-training.
- The SIM model employs novel pre-training tasks, such as masked bar-attribute modeling and score-MIDI matching, to capture music structures and align visual and symbolic representations.

## Methodology critique

- The authors employ a pre-trained Vision Transformer (ViT) to initialize the interaction transformer weights, leveraging the transformer's inherent ability to process visual features without the need for a separate visual embedder.
- The use of a single-stream approach that streamlines the processing of both visual and symbolic musical data is a deviation from conventional practices.

# Article 6 Strengths and Weaknesses

Strengths	Weaknesses
<ul style="list-style-type: none"><li>- The integration of score images as a distinct modality within the multimodal pre-training framework, which enriches the model's comprehension of symbolic music.</li><li>- The use of novel pre-training tasks that enable the SIM model to capture music structures and align visual and symbolic representations.</li><li>- The curation of a specialized dual-modality dataset that optimizes SIM model training.</li></ul>	<ul style="list-style-type: none"><li>- The model's performance is not evaluated on a diverse range of musical genres and notational styles, which may limit its applicability.</li><li>- The computational efficiency of the model is not thoroughly examined, which may be a concern for larger datasets and more complex musical compositions.</li><li>- The authors do not explore additional pre-training tasks that could further enhance the model's ability to capture the nuances of musical expression.</li></ul>

A decorative graphic of musical notation, including a treble clef, a staff with a key signature of one sharp (F#), and various musical notes and rests, positioned in the bottom right corner of the page.

Literature Review

# Extending NLP Transformers for Structured Music Composition

---

By David Daniel, Danil Afonchikov, Ayhem Bouabid