Comparison of symbolic transformer neural network models for analysis of phrase structure in Mozart piano sonatas

Ruan Buitendag (u20427222@tuks.co.za), Prof. J. Pieter Jacobs (jpjacobs@up.ac.za)

Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa

Phrase structure analysis is a core component of music structure analysis in Western classical music. While earlier research has focused on the automatic prediction of phrase beginnings and endings for monophonic music, at present there is a paucity of studies involving music based on homophonic and/or polyphonic textures, i.e. the bulk of the standard repertoire. Our study presents a comparative analysis of transformer neural network models for phrase structure prediction in Mozart piano sonatas. Compared to older recurrent neural networks such as long short-term memory (LSTM) networks, transformers implement attention mechanisms that enable greatly improved context awareness, often enhancing predictions for longer sequences.

We compare two distinct symbolic input representations: sequences of chords represented in Roman numeral notation, and sequences of the actual notes comprising the music. We note that harmonic structure, along with the melodic surface, are integral to perceived form. For transformers, processing time scales quadratically with token sequence length. Sequences based on Roman numeral annotations are typically far shorter than the corresponding note sequences, and therefore less expensive to process.

We evaluate the accuracy of model predictions of the phrase beginnings and endings annotated in the Batik-Mozart dataset. We present global classification metrics (precision, recall and F1 score), and targeted analyses of individual music examples. Overall, the note-level model (macro-average F1-score: 0.58), which is based on MusicBERT, somewhat outperforms the chord-level model. As expected, the chord-level model performs well in harmonically dense sections. In slow movements, the note-level model at times appears to be hampered by embellishing note flourishes superimposed on a regular harmonic rhythm. Interestingly, false-positive phrase boundaries predicted by either model often happens at musically plausible instances.

We discuss the two methods' possible complementarity and describe a musically informed data augmentation strategy that improved the chord-level model's generalization.

We would prefer to present orally but are also happy to do a poster presentation.