MIDI Transformer Tokenization

Jonas Veit, NAMEN EINFÜGEN

January 7, 2025

miditok

- Specify the data
- train the tokenizer (Do we need this? I think yes for efficiency)
- use Byte Pair Encoding (BPE)
- save the tokenizer and vocabulary
- NOTE: our goePT expects integers as tokens

Different tokenizers

- For simple melodies: REMI or MelodyTokenizer.
- For more complex, polyphonic data: Consider PolyphonyTokenizer or NoteTokenizer.
- For structured MIDI data: Use Structured or TSD.
- For music generation: REMI or MIDI-Like might work well.