An Exploration of Generating Sheet Music Images

Marcos Acosta[1] Irmak Bükey[2] TJ Tsai[1]
[1] Harvey Mudd College [2] Pomona College

HARVEY MUDD COLLEGE

ISMIR 2022 BENGALURU

Problem Statement

Goal: Explore and evaluate 5 different methods of generating sheet music images.

Why generate sheet music?

- In many genres, it is the main format that musicians use to learn a composition
- There is a lot more sheet music data than MIDI

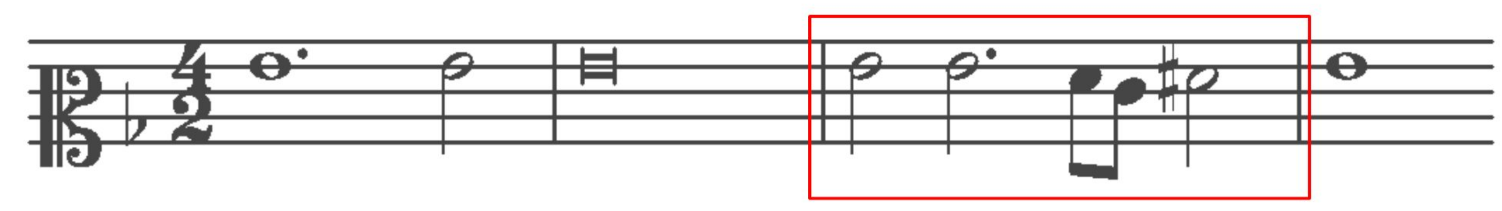
Seed

Sheet Music Generator

Continuation

Dataset

PrIMuS dataset [1] contains >8k short excerpts (incipits) in the following formats:



(a) Image format (png)

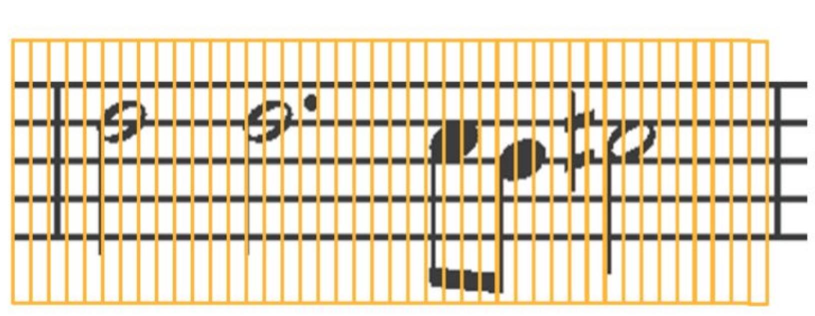
(b) Semantic encoding of musical symbols ("semantic representation")

```
clef-C2 keySignature-FM timeSignature-4/2
note-G4_whole.  note-G4_half    barline
note-G4_double_whole    barline note-G4_half
note-G4_half.   note-F4_eighth  note-E4_eighth
note-F#4_half   barline note-G4_whole
```

(c) XML representation of sheet music (MEI)

```
<layer xml:id="layer-0000000270816441" n="1">
  <note xml:id="note-0000001094075894" dur="2" oct="4" pname="g" />
  <note xml:id="note-0000001378564844" dots="1" dur="2" oct="4" pname="g" />
  <beam xml:id="beam-0000000841145766">
    <note xml:id="note-0000000338265625" dur="8" oct="4" pname="f" />
    <note xml:id="note-0000000252040961" dur="8" oct="4" pname="e" />
  </beam>
  <note xml:id="note-0000001214679643" dur="2" oct="4" pname="f">
    <accid xml:id="accid-0000001141211519" accid="s" />
  </note>
</layer>
```

Generating PNGs Directly



Pixel Columns

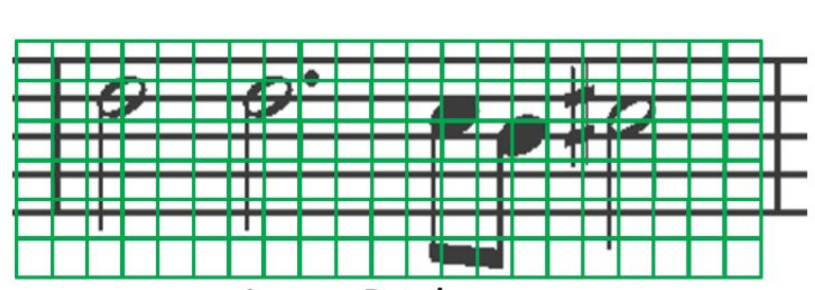
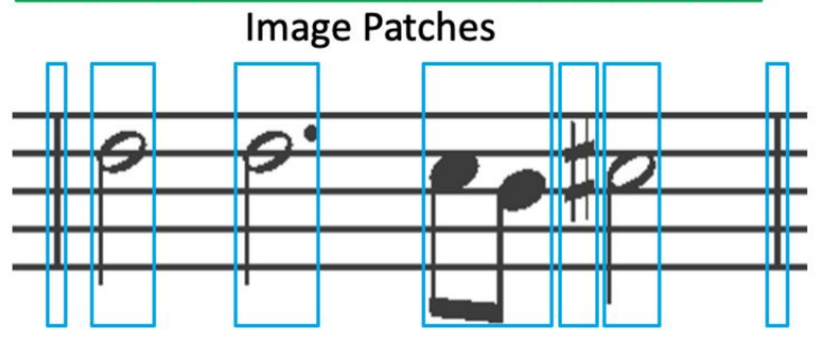


Image Patches



Visual Word Tokens

As an initial exploration, we generated images directly on a pixel-level. We experimented with three methods of dividing sheet music into tokens, and trained AWD-LSTM and GPT-2 models to predict the next token.

(1) Pixel columns: Predict a sequence of binary pixel columns

(2) Image patches: Predict a sequence of binary NxN image patches

(3) Visual word tokens: Predict a sequence of "whitespace"- separated visual tokens

Generating Visual Encodings

(4) Semantic Encoding Generation

Seed

`clef-G2`

Sample tokens with GPT/AWD-LSTM

Semantic encoding

`clef-G2 timeSignature-2/4 rest-eighth note-A4_eighth …`

Conversion algorithm

PAE encoding

`%G-2@3/4$bBEAü ''4B …`

Engraver (Verovio)

Sheet music image

Engraver (Verovio)

Instead of generating pixels directly, we also trained language models on two symbolic representations of sheet music.

Since the semantic encoding is unique to PrIMuS, it is converted to a more standard encoding before engraving.

(5) Sheet Music XML (MEI) Generation

Seed

`<music>`

GPT / AWD-LSTM with sampling

Semantic encoding

`<music> <body> <mdiv> <scoreDef meter.count="2" meter.unit="4"> …`

Qualitative Results

Pixel columns
Contains defects, especially in beams. Poor rhythmic cohesion

Image patches
Often fails to find right "patch" for flags and beams (long-range visual symbols)

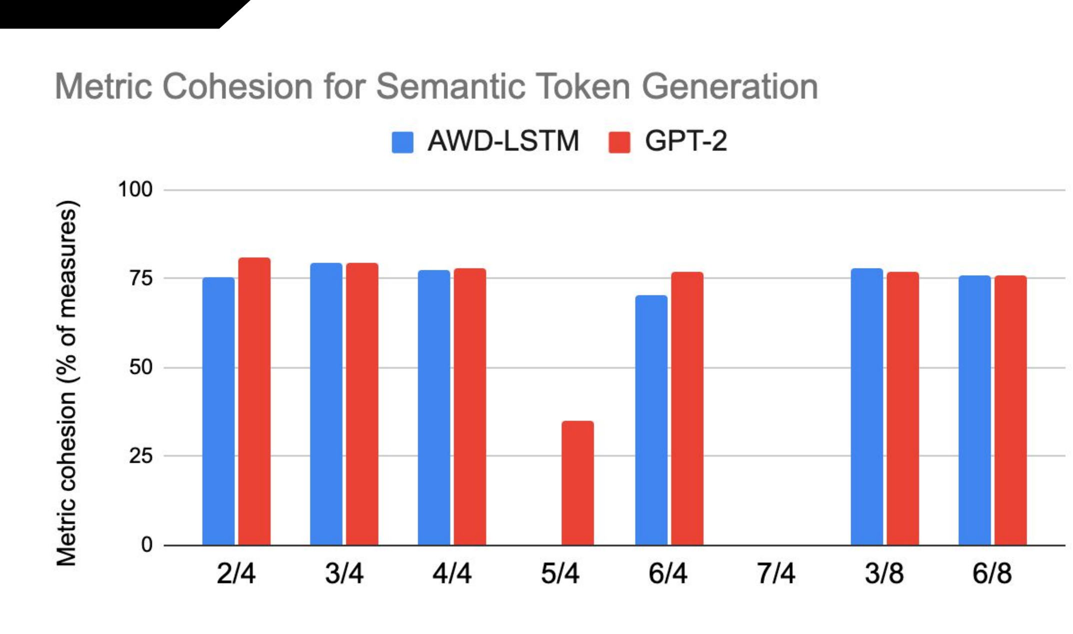
Visual Word Tokens
Looks visually better, but does not obey stylistic conventions; rhythmic incoherence

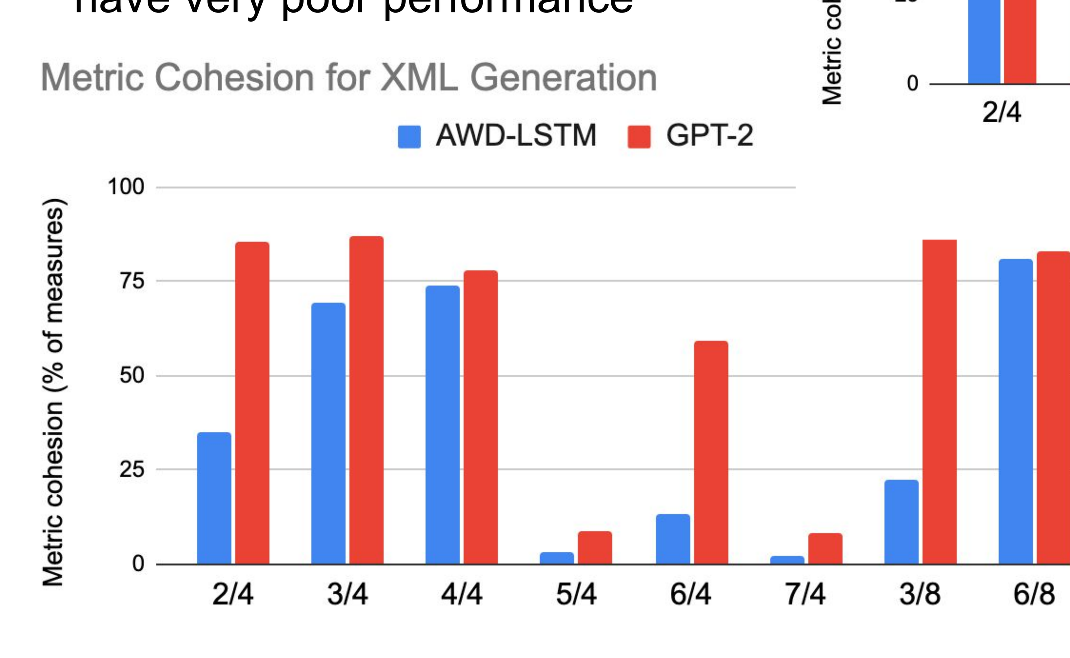
Semantic Encoding
Visually cohesive; more metrically coherent than pixel-based methods but still flawed in many cases

MEI (XML)
Visually similar to semantic encoding, metric coherence is the main issue

Metric Cohesion Analysis

- Bars represent percent of generated measures that have the correct number of beats
- Metric cohesion is roughly equal for both models with semantic token generation
- Rare time signatures (5/4, 7/4) have very poor performance





- Performance on XML (MEI) generation is much more varied
- With XML, GPT outperforms LSTM on all time signatures, sometimes significantly
- Metric cohesion is still low enough to be prohibitive; future work must encode rhythmic properties into the model

References & Acknowledgements

[1] Calvo-Zaragoza, J.; Rizo, D. *End-to-End Neural Optical Music Recognition of Monophonic Scores*. Appl. Sci. 2018, 8, 606

This material is based upon work supported by the National Science Foundation under Grant No. 2144050.