Multitrack Music Transformer

Hao-Wen Dong Ke Chen Shlomo Dubnov Julian McAuley Taylor Berg-Kirkpatrick

University of California San Diego





UC San Diego

Overview

Generate orchestral music

- of diverse instruments
- using a new compact representation
- with a multi-dimensional transformer

(Source: Vienna Mozart Orchestra)

Demo

Related Work (Transformers for Music Generation)

Model	Multitrack	Instrument control	Compound tokens	Generative modeling
REMI [5]				✓
MMM [10]	✓			✓
CP [6]			✓	✓
MusicBERT [15]	✓		✓	
FIGARO [11]	✓			✓
MMT (ours)	✓	✓	✓	✓

	Average sample length (sec)	Inference speed (notes per second)
MMM [10]	38.69	5.66
REMI+ [11]	28.69	3.58
MMT (ours)	**100.42**	**11.79**

→ Longer samples!
Faster inference speed!

Huang and Yang, "Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions," *Proc. MM*, 2020.
Ens and Pasquier, "MMM : Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008.06048*, 2020.
Hsiao et al., "Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs," *Proc. AAAI*, 2023.
Zeng et al., "MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training," *Proc. Findings of ACL*, 2021.
von Rütte et al., "FIGARO: Controllable Music Generation using Learned and Expert Features," *Proc. ICLR*, 2023.

Representation

- We represent a music piece as a sequence of events

$$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

- Each event $\mathbf{x}_i$ is encoded as

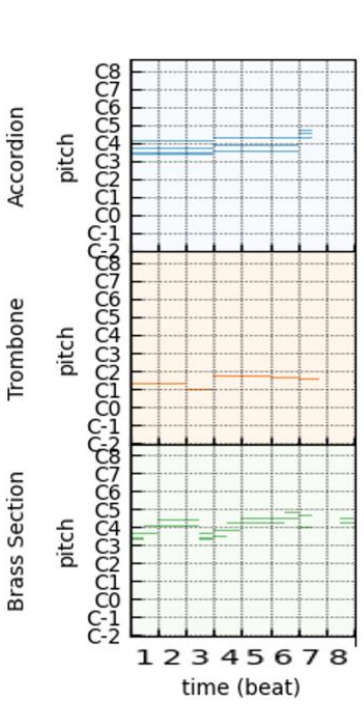
$$\mathbf{x}_i = (x_i^{type}, x_i^{beat}, x_i^{position}, x_i^{pitch}, x_i^{duration}, x_i^{instrument})$$

Specify note & instrument information

5 event types

- Structural
  - Start of song
  - Start of notes
  - End of song
- Data
  - Instrument
  - Note

Representation (An Example)



Structural events

```
(0,  0,  0,  0,  0,  0)   Start of song
(1,  0,  0,  0,  0, 15)   Instrument: accordion
(1,  0,  0,  0,  0, 36)   Instrument: trombone
(1,  0,  0,  0,  0, 39)   Instrument: brasses
(2,  0,  0,  0,  0,  0)   Start of notes
(3,  1,  1, 41, 15, 36)   Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3,  1,  1, 65,  4, 39)   Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3,  1,  1, 65, 17, 15)   Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3,  1,  1, 68,  4, 39)   Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3,  1,  1, 68, 17, 15)   Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3,  1,  1, 73, 17, 15)   Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
(3,  1, 13, 68,  4, 39)   Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
(3,  1, 13, 73,  4, 39)   Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
(3,  2,  1, 73, 12, 39)   Note: beat=2, position=1,  pitch=C5, duration=36, instrument=brasses
(3,  2,  1, 77, 12, 39)   Note: beat=2, position=1,  pitch=E5, duration=36, instrument=brasses
        ...                       ...
(4,  0,  0,  0,  0,  0)   End of song
```

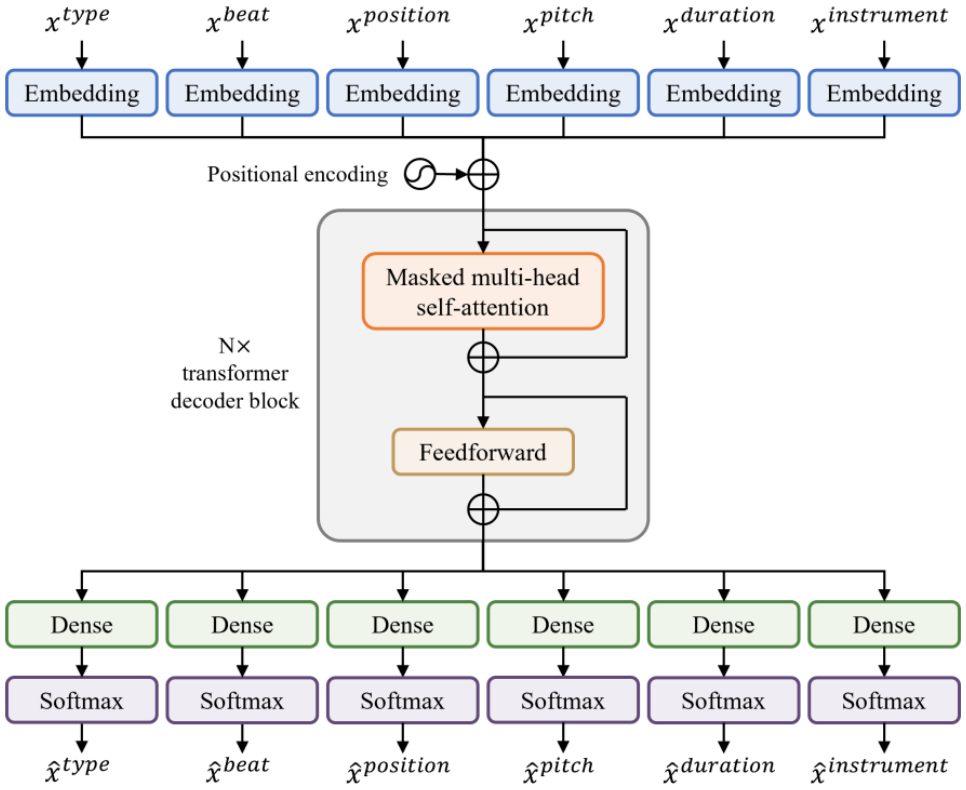
Instrument events

Note events

Hierarchically sorted!

Multitrack Music Transformer

- A multi-dimensional decoder-only transformer model
  - Predict six fields *at the same time*
- Trained autoregressively
  - Predict the next event given past events
- At inference time, illegal values are assigned zero probabilities
  - Violate the ordering of structural events
  - Violate the hierarchical sorting of events



Three Sampling Modes

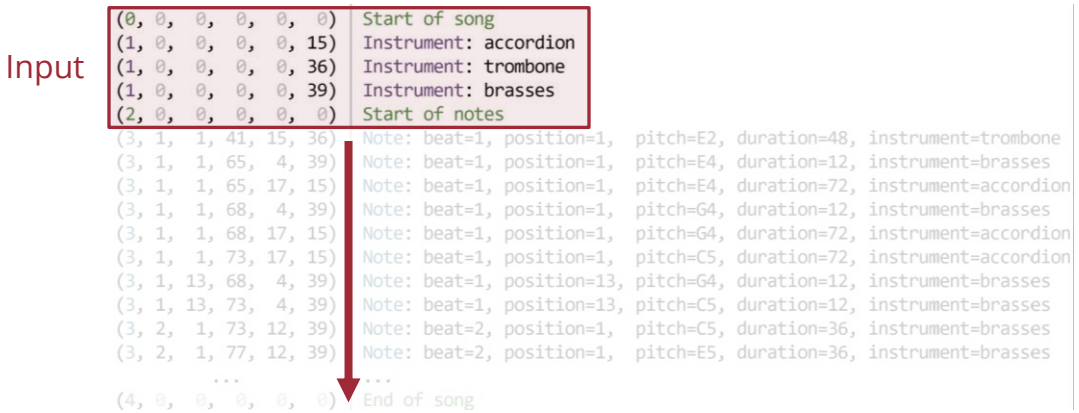
Unconditional generation

Input

Only need to train ONE model!

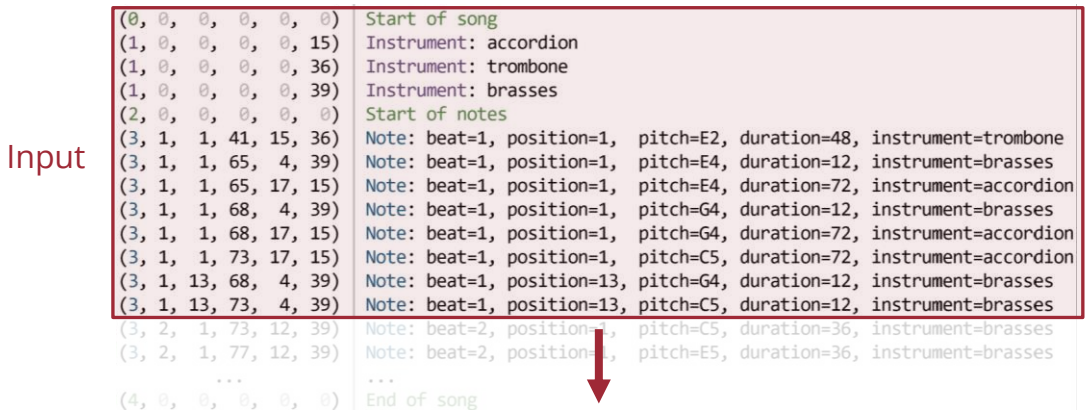
Instrument-informed generation

Input



N-beat continuation

Input



Experimental Setup

Data

- Symbolic Orchestral Database (SOD) (Crestel et al., 2017)
  - 5,743 songs, 357 hours
- Temporal resolution: 12 time steps per quarter note
- 80% training, 10% validation, 10% test
- Data augmentation
  - Randomly shift for -5~6 semitones
  - Randomly select a starting beat

Model & Training

- 6 transformer decoder blocks
- 8 attention heads
- Model dimension: 512
- Sequence length: 1024
- Maximum number of beats: 256
- Maximum training steps: 200,000

Crestel et al., "A database linking piano and orchestral MIDI scores with application to automatic projective orchestration," *Proc. ISMIR*, 2017.

Example Results

Unconditional generation

Instrument-informed generation

church-organ, viola, contrabass, strings, voices, horn, oboe

4-beat continuation

Wolfgang Amadeus Mozart's Eine kleine Nachtmusik

More audio samples

salu133445.github.io/mmt/

OttoPilot33, "Eine Kleine Nachtmusik - By My MIDI Virtual Orchestra." https://youtu.be/_g8iCdY34jc

Subjective Listening Test Results

	Number of parameters	Average sample length (sec)	Inference speed (notes per second)	Subjective listening test results			
				Coherence	Richness	Arrangement	Overall
MMM [10]	19.81 M	38.69	5.66	3.48 ± 0.35	3.05 ± 0.38	3.28 ± 0.37	3.17 ± 0.43
REMI+ [11]	20.72 M	28.69	3.58	$\mathbf{3.90 \pm 0.52}$	$\mathbf{3.74 \pm 0.21}$	$\mathbf{3.74 \pm 0.44}$	$\mathbf{3.77 \pm 0.41}$
MMT (ours)	19.94 M	**100.42**	**11.79**	3.55 ± 0.46	3.53 ± 0.35	3.40 ± 0.44	3.33 ± 0.47

2.6x/3.5x longer generated samples
(within the same sequence length)

2.1x/3.3x faster inference speed

Higher quality than MMM
Lower quality than REMI+

Ens and Pasquier, "MMM : Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008.06048*, 2020.
von Rütte et al., "FIGARO: Controllable Music Generation using Learned and Expert Features," *Proc. ICLR*, 2023.

Analyzing Self-attention

- *Mean relative attention* for a field d:

$$\gamma_k^{(d)} = \frac{\sum_{x \in \mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x}) \, 1_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{x \in \mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})}$$

Attention weight

Whether the field value is of difference k

$\gamma_{-8}^{(pitch)}$

```
(0, 0,  0,  0,  0,  0)   Start of song
(1, 0,  0,  0,  0, 15)   Instrument: accordion
(1, 0,  0,  0,  0, 36)   Instrument: trombone
(1, 0,  0,  0,  0, 39)   Instrument: brasses
(2, 0,  0,  0,  0,  0)   Start of notes
(3, 1,  1, 41, 15, 36)   Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3, 1,  1, 65,  4, 39)   Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3, 1,  1, 65, 17, 15)   Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3, 1,  1, 68,  4, 39)   Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3, 1,  1, 68, 17, 15)   Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3, 1,  1, 73, 17, 15)   Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
```

$\gamma_{-5}^{(pitch)}$

Analyzing Self-attention

- *Mean relative attention* for a field d:

$$\gamma_k^{(d)} = \frac{\sum_{x \in \mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})\, 1_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{x \in \mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})}$$

Biased towards difference that occurred more frequently!

- *Mean relative attention gain* for a field d:

$$\tilde{\gamma}_k^{(d)} = \gamma_k^{(d)} - \frac{\sum_{x \in \mathcal{D}} \sum_{s>t} 1_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{x \in \mathcal{D}} \sum_{s>t} \boxed{1}}$$

Assuming a uniform attention matrix

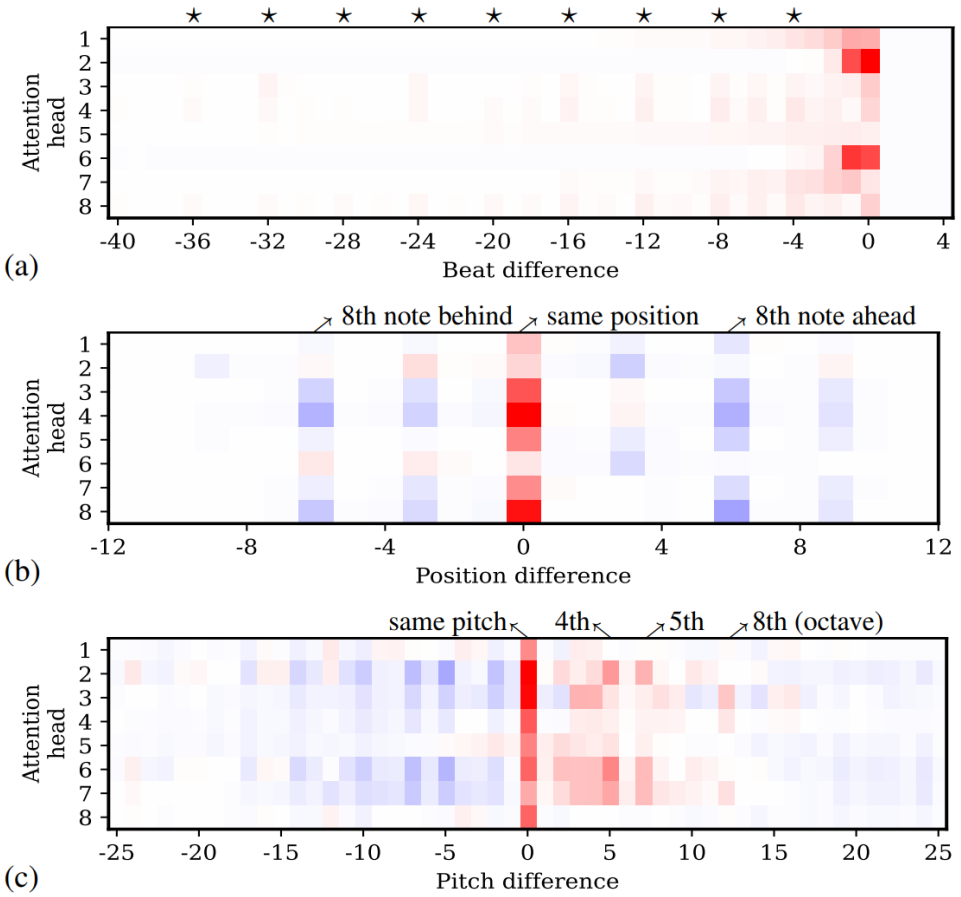
Musical Self-attention

The MMT model attends more to notes

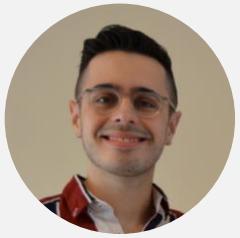
- that are $4N$ beats away in the past
- that have the same position as the current note (A note on beat attends more to a note on beat; a note off beat attends more to a note off beat.)
- that has a pitch in an octave above which forms a consonant interval

→ MMT learns a relative self-attention for certain aspects of music, specifically, beat, position and pitch.

Positive and negative mean relative attention gain



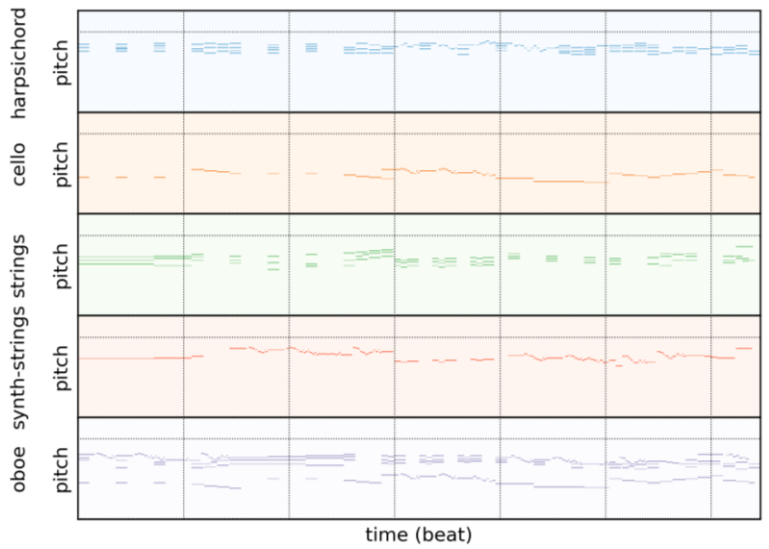
Extension: Music Reduction



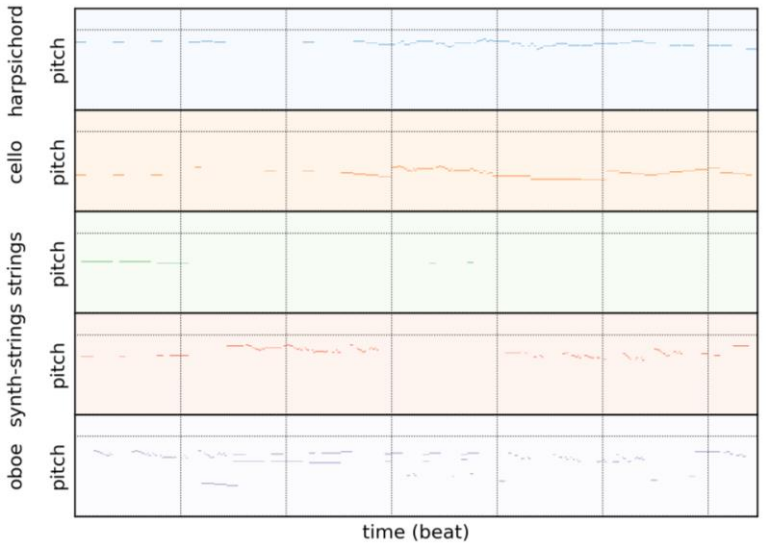
Zachary Novack
UC San Diego
znovack@ucsd.edu
zacharynovack.github.io

- Using the proposed representation, we can build a music reduction system that **simplifies music while keeping its *core elements***
- Could be applied to controlled music rearrangement for **music education**

Input



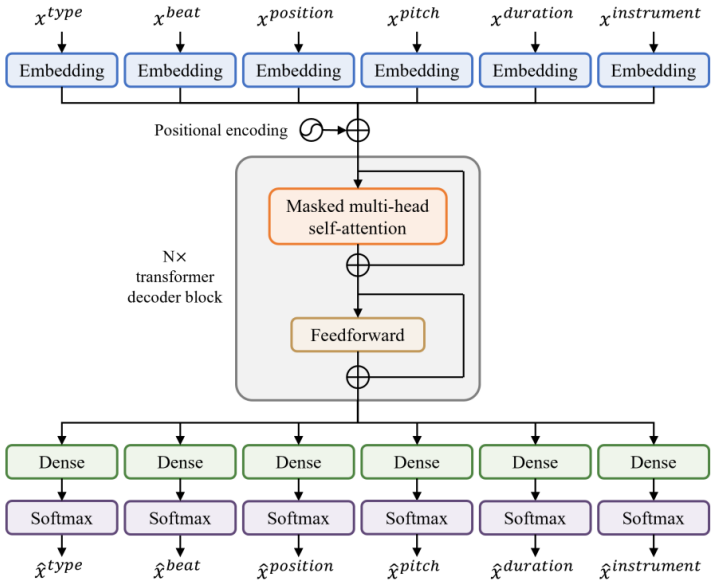
Output



Summary

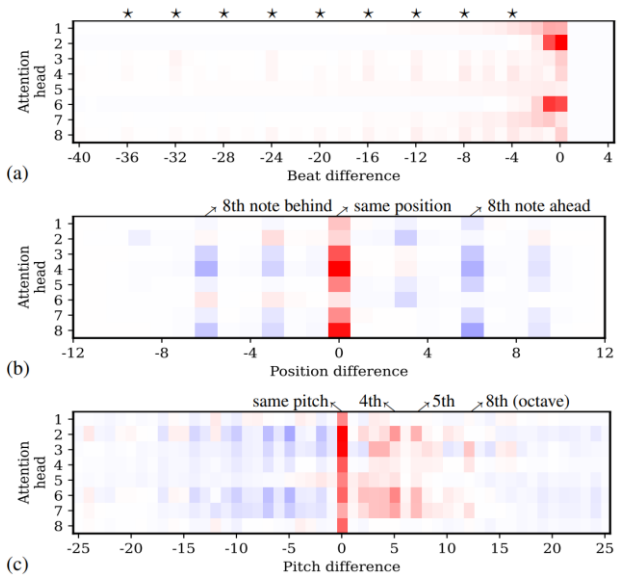
Multitrack Music Transformer

Proposed an efficient representation and model for multitrack music generation



Musical Self-attention

Presented the first systematic analysis of musical self-attention

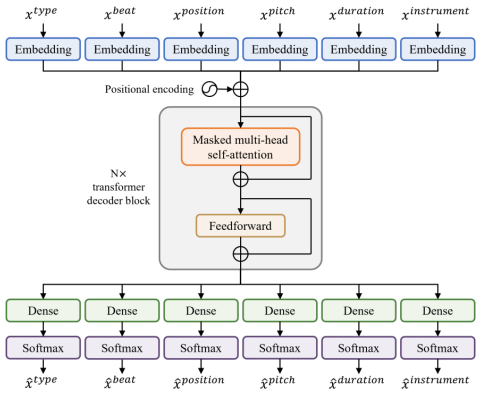


Acknowledgements

- Hao-Wen thanks J. Yang and Family Foundation and Taiwan Ministry of Education for supporting his PhD study.
- This project has received funding from the European Research Council (ERC REACH) under the European Union's Horizon 2020 research and innovation programme (Grant agreement #883313).

Thank you!

Multitrack Music Transformer

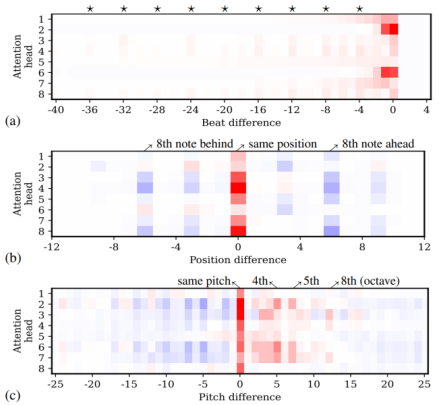


Paper
arxiv.org/abs/2207.06983

Demo
salu133445.github.io/mmt/

Code
github.com/salu133445/mmt

Musical Self-attention







UC San Diego