SymphonyNet

Symphony Generation with Permutation Invariant Language Model

Demo/Code/Dataset

Jiafeng Liu[1], Yuanliang Dong[1], Zehua Cheng[2], Xinran Zhang[1], Xiaobing Li[1], Feng Yu[1], Maosong Sun[1,3]
[1] Central Conservatory of Music [2] University of Oxford [3] Tsinghua University
{jiafeng.liu, gunterdong}@mail.ccom.edu.cn

Introduction

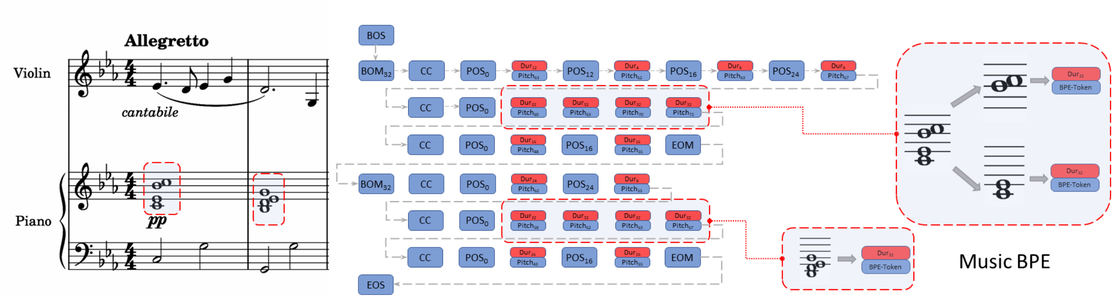
Symphony Generation:
Generating multi-track symbolic music with repeatable instruments.

Permutation Invariance:
Notes playing simultaneously have no priori order.

Challenges:
- Flatten a 2-D score into sequential tokens and distinguish tracks with the same instrument.
- Build a language model where simultaneous notes could be swapped.
- Generate symphony music which becomes an ultra-long sequence of tokens after flattening.

Solutions:
- Multi-track Multi-instrument Repeatable (MMR) representation
- 3-D Positional Embedding
- Music BPE and a linear transformer decoder

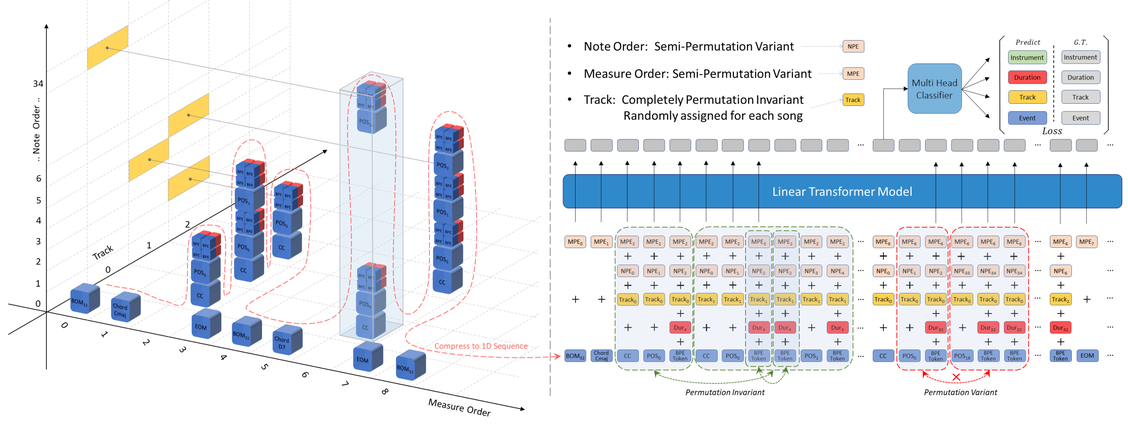
MMR Representation with Music BPE



MMR Representation:
- Beat-based time-unit suitable for score notes
- Align different tracks by Measure
- Structural and controlling tokens, like:
  - $[BOS]$, $[EOS]$, $[BOM_i]$, $[EOM]$ indicates a score or a measure's beginning or ending
  - $[CC]$ indicates a beginning of another track within the same measure
  - $[POS_i]$, $[Chord_i]$, etc.
- Note attributes: event, duration, track, instrument

Music BPE:
- A trade-off between *Pitch* ("character-level") and *Chord* ("word-level"): Find "Subword" in music.
- Difference: Music BPE is based on concurrence of notes rather than adjacency of characters in the original BPE algorithm.

3-D Positional Embedding



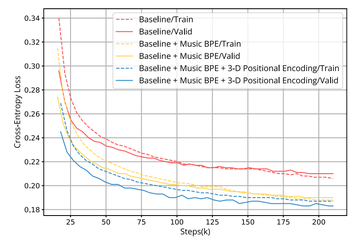
Good News! Self-attention modules in Transformer are designed premutation invariant.

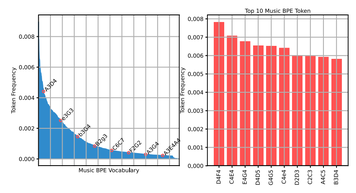
Use 3-D Positional Embedding to tell the model:
- **Simultaneous Notes within the same track** have totally the same positional embedding.
- **Simultaneous Notes in different tracks** just have different "track labels", learned by randomly assigned track numbers.
- **Successive notes within the same track** are located by their unique measure position embedding (MPE) and note position embedding (NPE).

Experiments and Results

Experiments:
- Collected Symphony Dataset
  - 46,359 symphony scores with 279M notes
- Ablation Study (on Symphony Dataset):
- Comparison to MMM[1] on LMD Dataset[2]
- Subjective Listening Test
  - 50 participants including 25 musicians
  - Rating in 5 dimentions: Coherence, Diversity, Harmoniousness, Structureness, Orchestration and overall Preference

Results:
- Objective results
- Music BPE "subword" analysis
- Subjective results:
  - Music BPE + 3-D positional embedding improves generation quality.
  - SymphonyNet surpasses MMM in all indicators and could construct coherent, unique, complex, and harmonic symphonies.





	Model	C	D	H	S	O	P
Chord	Baseline	3.5	3.57	3.07	3.00	3.21	3.29
	BPE	3.64	3.64	3.14	**3.15**	3.43	3.29
	3D + BPE	**3.71**	**3.72**	**3.21**	3.07	**3.5**	**3.5**
	Human	4.43	3.43	4.14	4.36	4.14	4.14
Prime	Baseline	3.79	2.79	3.21	3.43	3.36	3.36
	BPE	**3.86**	**3.5**	**3.5**	3.5	3.64	**3.86**
	3D + BPE	**3.86**	3.14	3.43	**3.57**	**3.93**	3.64
	Human	4.36	3.57	4.36	4.00	4.36	4.36
Uncondi.	Baseline	3.52	3.46	3.04	3.07	3.11	3.07
	BPE	**3.79**	3.64	3.25	3.11	3.25	3.29
	3D + BPE	3.53	**3.93**	**3.43**	**3.32**	**3.43**	**3.32**
	Human	4.39	3.89	4.18	4.21	4.11	4.29

(a) Trained on Symphony Dataset

Model	C	D	H	S	O	P
MMM	3.20	2.71	2.51	2.66	2.80	2.71
Symph.	**3.33**	**2.89**	**2.76**	**2.69**	**2.99**	**2.87**

(b) Trained on Lakh MIDI Dataset

[1] J. Ens and P. Pasquier, "Mmm: Exploring conditional multi-track music generation with the transformer," arXiv preprint arXiv:2008.06048, 2020.
[2] C. Raffel, "Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching," Ph.D. dissertation, COLUMBIA UNIVERSITY, 2016.