Compound Word Transformer:

Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs

Wen-Yi Hsiao[1], Jen-Yu Liu[1], Yin-Cheng Yeh[1], Yi-Hsuan Yang[1,2]

[1]Taiwan AI Labs, Taipei, Taiwan
[2]Academia Sinica, Taipei, Taiwan

Motivation

- Transformer as a strong music generation model
  - Pop music
    - mean: 6432 tokens
    - **max >= 10K** tokens
  - memory complexity
    - vanilla transformer: $O(N^2)$
- Crop one song into segments
- Can we generate a music piece of full-song length?

Motivation

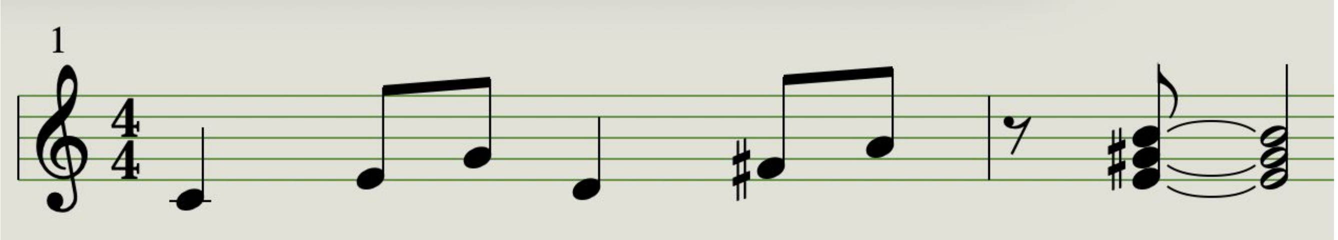
- Reduce the length of the token sequence
  - Novel compact representation of music
    - Compound Word (**CP**)

- Lower Memory Complexity
  - Advanced transformer
    - transformer-XL
    - linear transformer

Overview

- Representation: From MIDI to CP
- Model
- Experiments
    - Tasks
        - Unonditional Generation
        - Conditional Generation
    - Evaluation
        - Quantitative Evaluation
        - Qualitative Evaluation

Representation: MIDI



...
<note_on, note=60, velocity=94, time=0.0 sec>
<note_on, note=60, velocity=0, time=1.253 sec>
<note_on, note=64, velocity=94, time=1.253 sec>
<note_on, note=64, velocity=0, time=1.879 sec>
<note_on, note=67, velocity=94, time=1.879 sec>
<note_on, note=67, velocity=0, time=2.506 sec>
...

Sequence Length: 18
9 (notes) x 2 (note on/off)

...

pitch: 60,
velocity:94,
time: 1.20,

...

Sequence Length: 54
9 (notes) x 2 (note on/off) x 3 (attributes)

Representation: MIDI

- Problem
  - 1 note, 2 seperated events
  - absolute timing (second)

- *Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions* (Yu-Siang Huang, Yi-Hsuan Yang)
  - **REMI** representation
    - note off -> dutation
    - absolute timing -> symbolic timing (beat)

Representation: REMI

- duration

```
...
<note_on, note=60, velocity=94, time=0.0 sec>
<note_on, note=60, velocity=0,   time=1.253 sec>
<note_on, note=64, velocity=94, time=1.253 sec>
<note_on, note=64, velocity=0,   time=1.879 sec>
<note_on, note=67, velocity=94, time=1.879 sec>
<note_on, note=67, velocity=0,   time=2.506 sec>
...
```

Sequence Length: 54
9 (notes) x 2 (note on/off) x 3 (attributes)

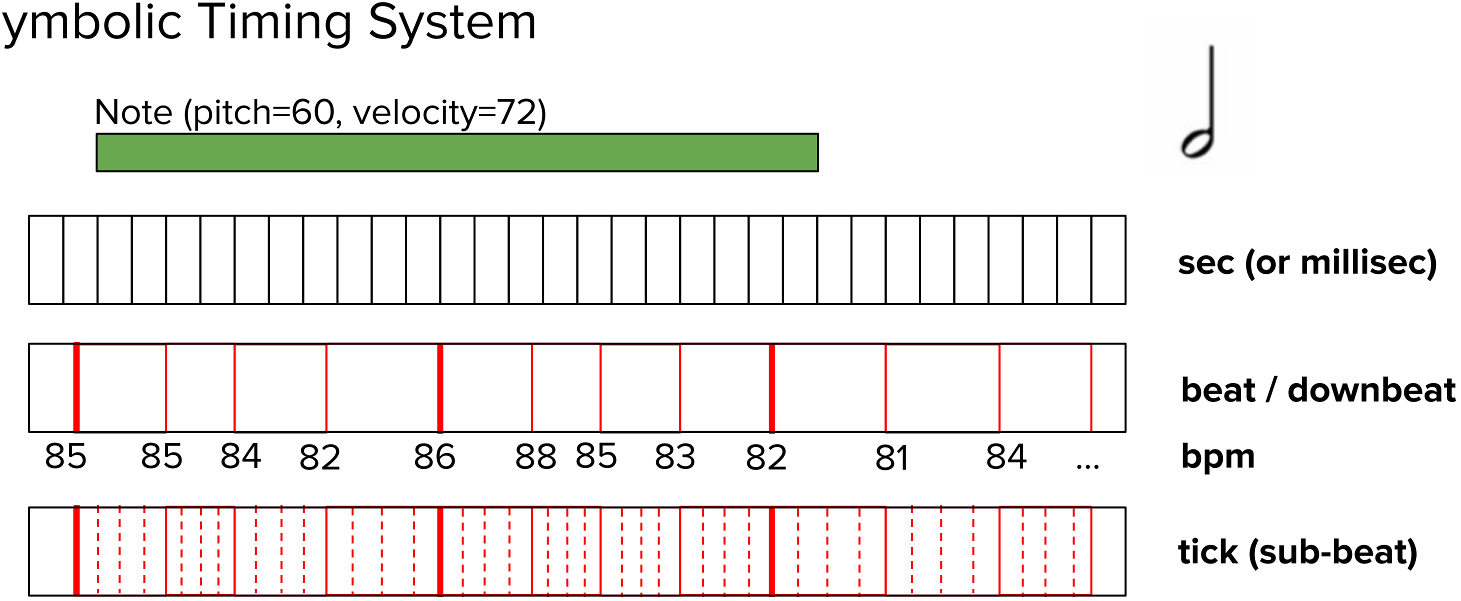
```
...
<note_on, note=60, duraiton=4, velocity=94, time=0 tick>
<note_on, note=64, duraiton=2, velocity=94, time=4 tick>
<note_on, note=67,  duraiton=2, velocity=94, time=6 tick>
...
```

Sequence Length: 36
9 (notes) x 4 (attributes)

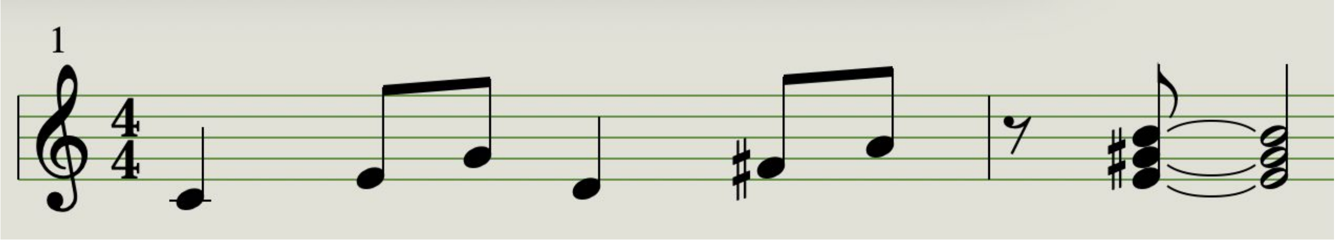
Representation: REMI

<note_on, note=60, duraiton=4 **tick**, velocity=94, time=0 **tick**>

- Symbolic Timing System



Representation: REMI



<bar>
<tempo, bpm=85, tme=0 tick>
<note_on, note=60, duraiton=4, velocity=94, time=0 tick>
<tempo, bpm=85, tme=4 tick>
<note_on, note=64, duraiton=2, velocity=94, time=4 tick>
<note_on, note=67, duraiton=2, velocity=94, time=6 tick>
...
<bar>
...

Sequence Length: 54

9 (notes) x 4 (attributes) +
8 (tempos) x 2 (attributes) +
2 (bars)

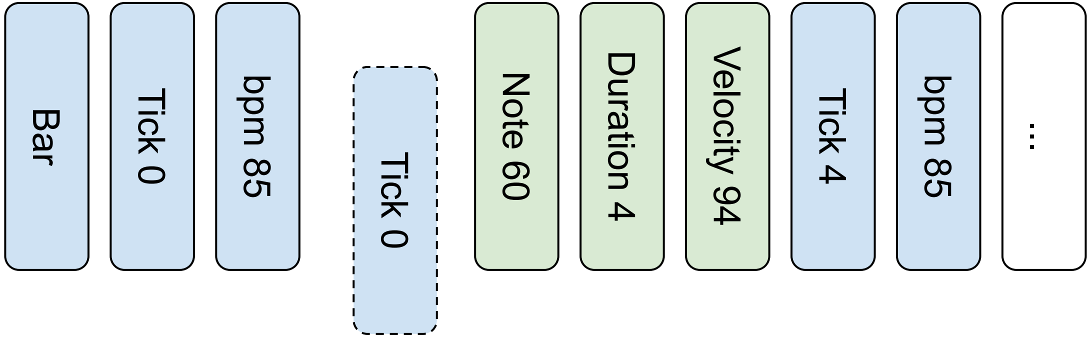
Representation: REMI v2

- Remove duplicated tokens

<bar>
<tempo, bpm=85, tme=0 tick>
<note_on, note=60, duraiton=4, velocity=94, time=0 tick>
<tempo, bpm=88, tme=4 tick>
...

Sequence Length: 50

9 (notes) x 4 (attributes) +
8 (tempos) x 2 (attributes) +
2 (bars) -
4 (duplicated events)



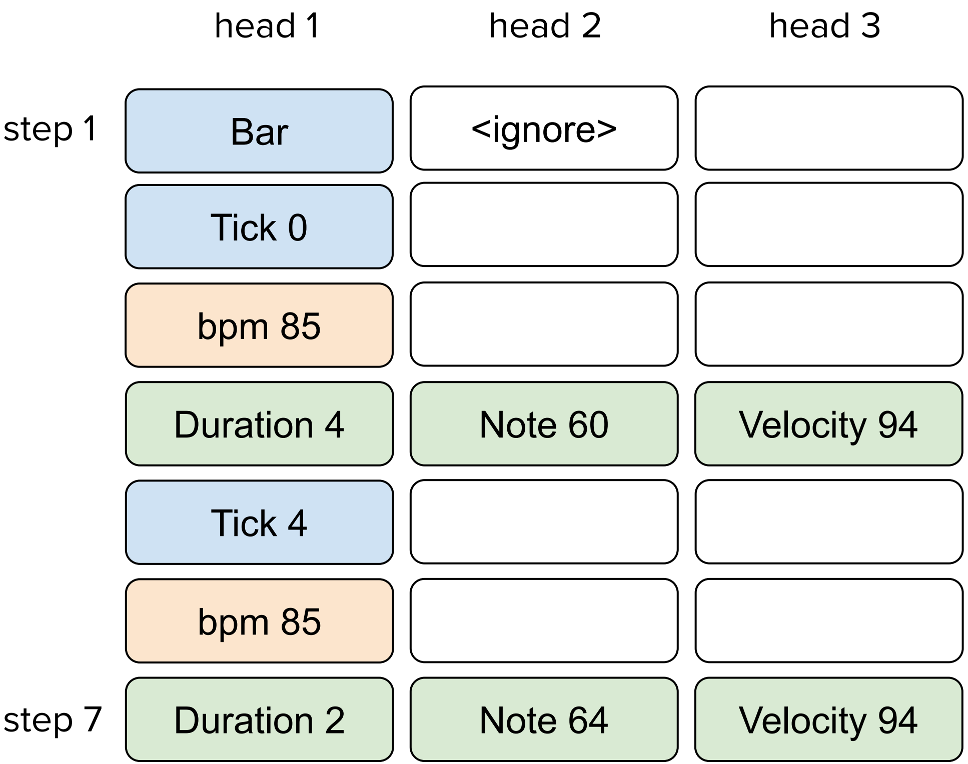
Remove

Representation: Compound Word (CP)

- E events, A attributes for each,
  - O(EA) steps in total
- **grouping**
- **multi-headed ouptut layer**

Sequence Length: 30

9 (notes) +
8 (tempos) +
2 (bars) +
11 (tick positions)



	head 1	head 2	head 3
step 1	Bar	<ignore>	
	Tick 0		
	bpm 85		
	Duration 4	Note 60	Velocity 94
	Tick 4		
	bpm 85		
step 7	Duration 2	Note 64	Velocity 94

Representation: CP

- expansion

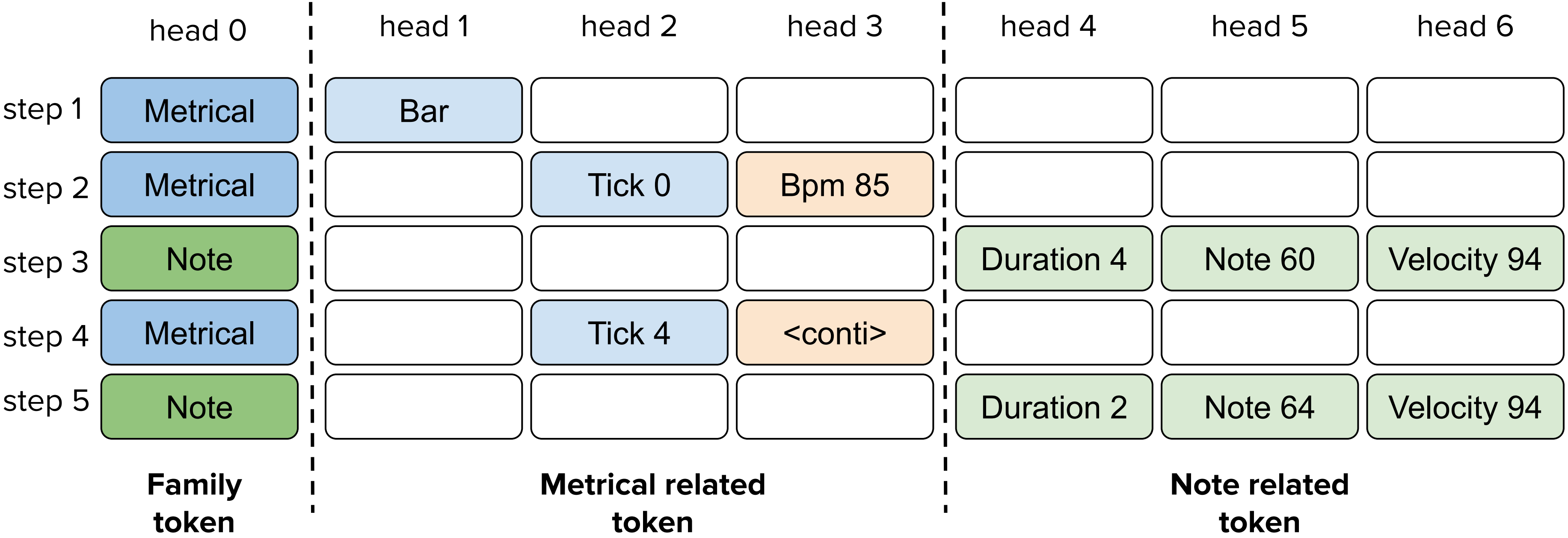
	head 1	head 2	head 3	head 4	head 5	head 6
step 1	Bar					
step 2		Tick 0				
step 3			Bpm 85			
step 4				Duration 4	Note 60	Velocity 94

Representation: Recap

- From MIDI to CP
  - Symbolize timing
  - Reduce about half sequence length

- The reduction in sequence length would be even greater when we add more attributes per events (e.g. chord)

Model: Multi-headed Ouput



	head 0	head 1	head 2	head 3	head 4	head 5	head 6
step 1	Metrical	Bar					
step 2	Metrical		Tick 0	Bpm 85			
step 3	Note				Duration 4	Note 60	Velocity 94
step 4	Metrical		Tick 4	<conti>			
step 5	Note				Duration 2	Note 64	Velocity 94
	Family token	**Metrical related token**			**Note related token**		

Model: Multi-headed Ouput

Stage 1

$$\mathbf{h}_t = \text{Self-attn}\left(\vec{\mathbf{x}}_{t-1}\right) ,$$

$$\widehat{f}_t = \text{Sample}_{\mathcal{F}}\left(\text{softmax}(\mathbf{W}_{\mathcal{F}}\mathbf{h}_t)\right) ,$$

Stage 2

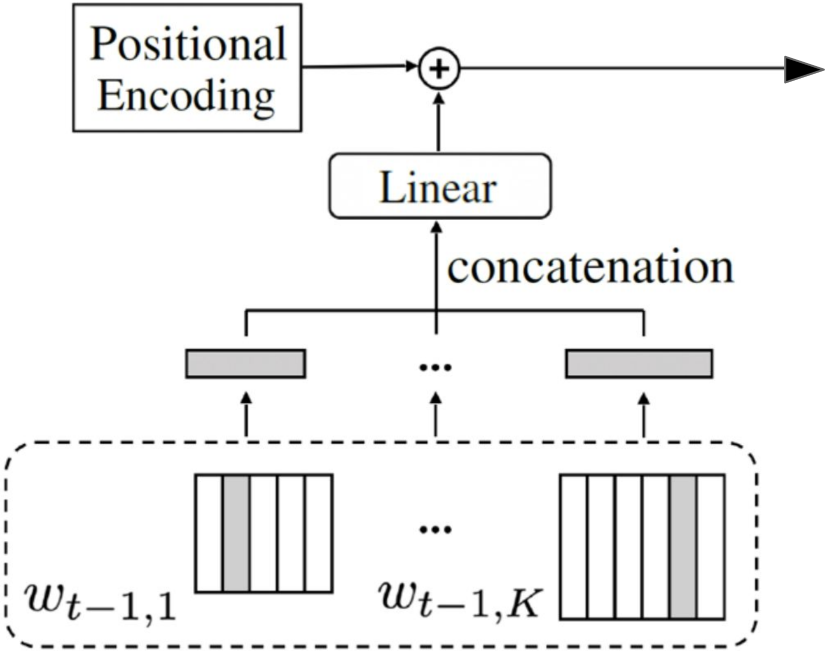
$$\mathbf{h}_t^{\text{out}} = \mathbf{W}_{\text{out}}\left[\mathbf{h}_t \oplus \text{Embedding}_{\mathcal{F}}(\widehat{f}_t)\right] ,$$

$$\widehat{w_{t,k}} = \text{Sample}_k\left(\text{softmax}(\mathbf{W}_k\mathbf{h}_t^{\text{out}})\right) , \; k = 1, ..., K ,$$

- Training: teacher forcing
- Inference: stochastic sampling

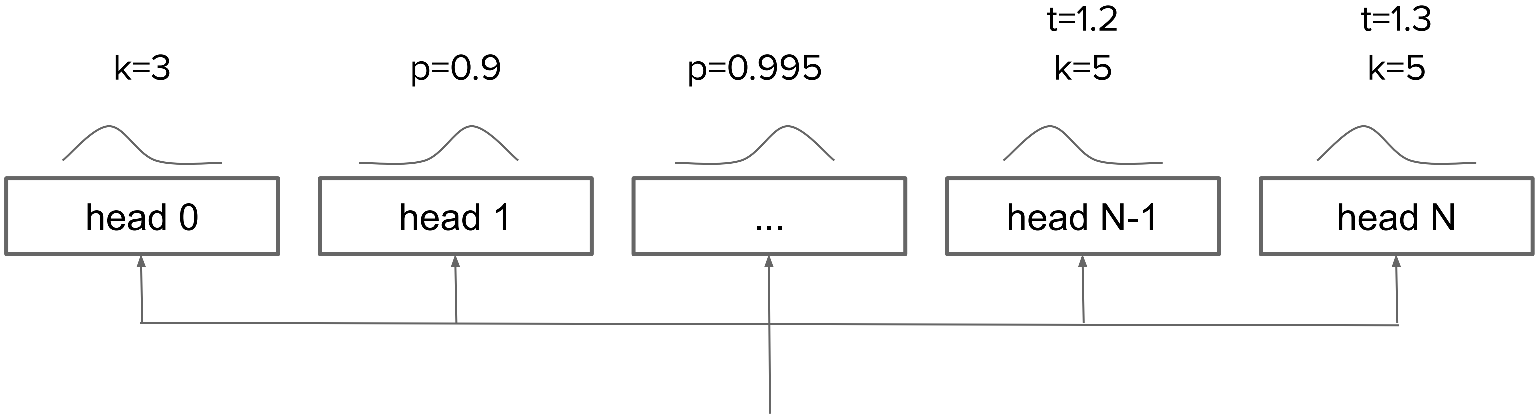
Model: Adaptive Embedding

- Difficulty of learning
    - Hard: velocity
    - Easy: bar & beat



Model: Adaptive Sampling

- Inference Stage
  - temperature t
  - top-k
  - top-p (nucleus)

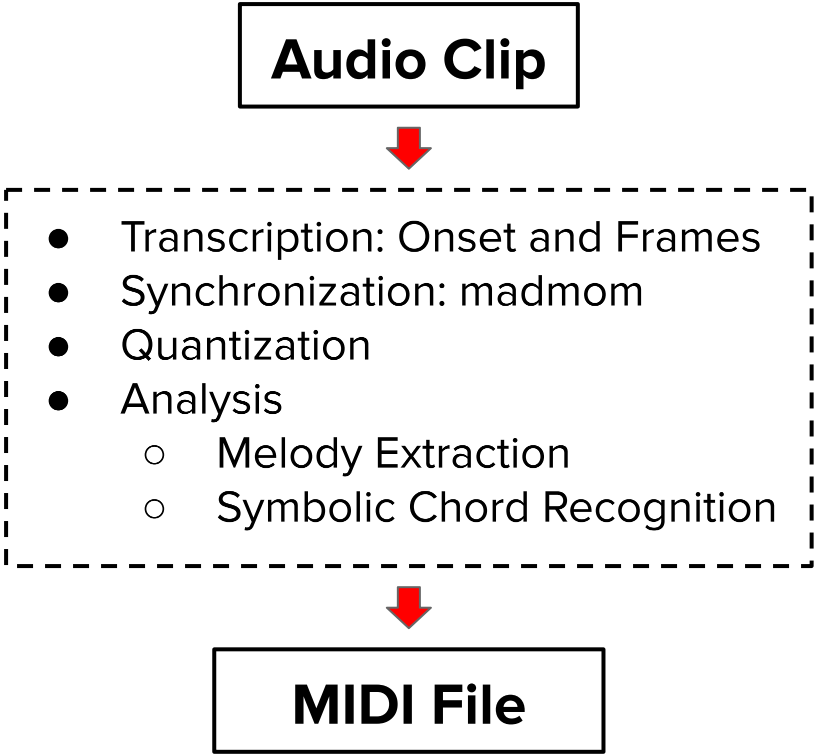


Model: Compound Word Transformer

- **Compound Word Transformer (CP Transformer)**
  - Compound Word Representation
  - Adaptive Embedding
  - Multi-Headed Ouput Module
  - Two-stage Prediction
  - Adaptive Sampling (Inference Time)

Dataset

- Corpus
  - 17K pop piano music dataset
- Tasks
  - Unconditional Generation
  - Conditional Generateion:
    - lead sheet to performance



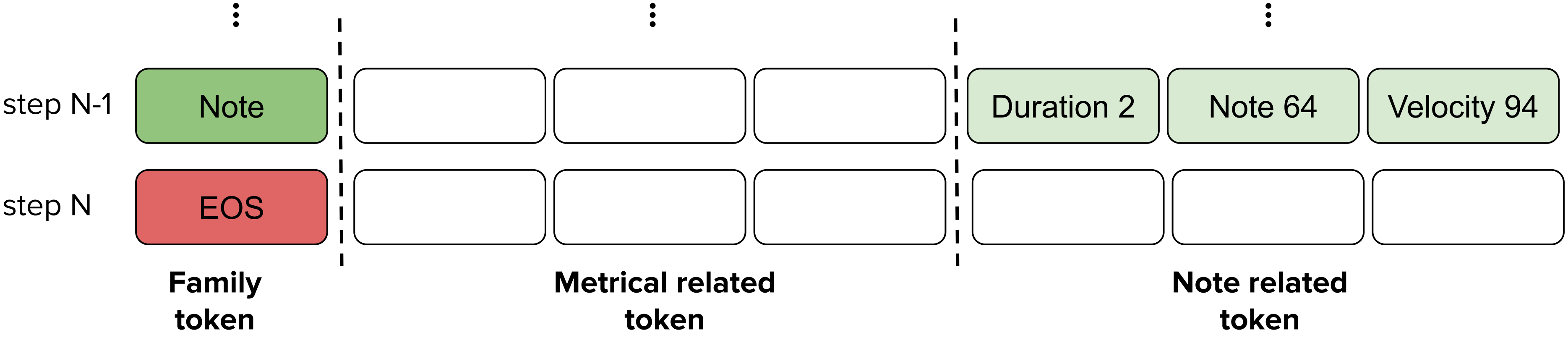
Audio Clip

- Transcription: Onset and Frames
- Synchronization: madmom
- Quantization
- Analysis
  - Melody Extraction
  - Symbolic Chord Recognition

MIDI File

Task1: Unconditional Generation

- Beat Sychronized Feature
  - Chord
  - Tempo
- New type token: **End of sequence** <EOS>



step N-1

Note

Duration 2

Note 64

Velocity 94

step N

EOS

Family token

Metrical related token

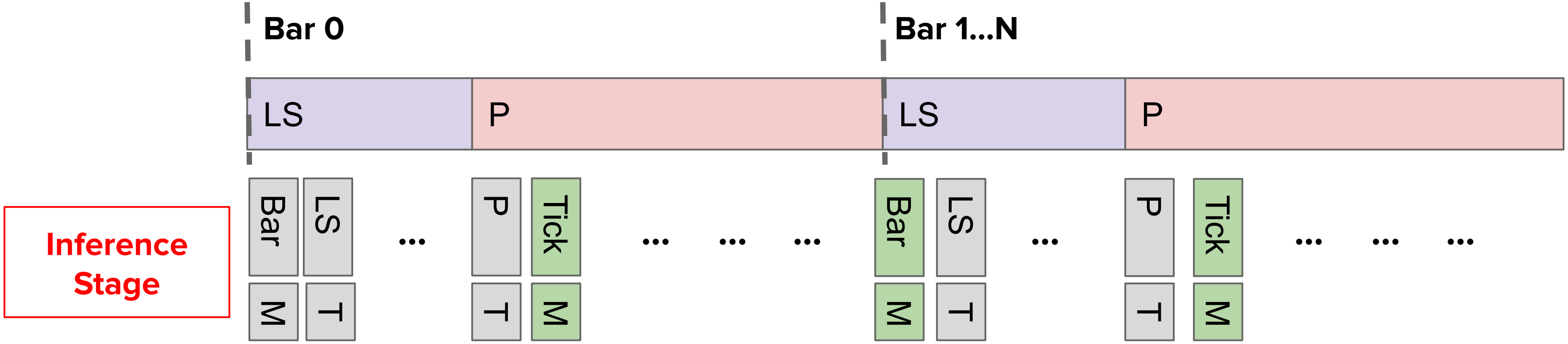
Note related token

Task2: Conditional Generation

- lead sheet to performance
- New type token: **track (T)**
  - **L**ead **S**heet track (**LS**)
  - **P**iano track (**P**)
- Encoder-free prefix LM (from google's "T5")

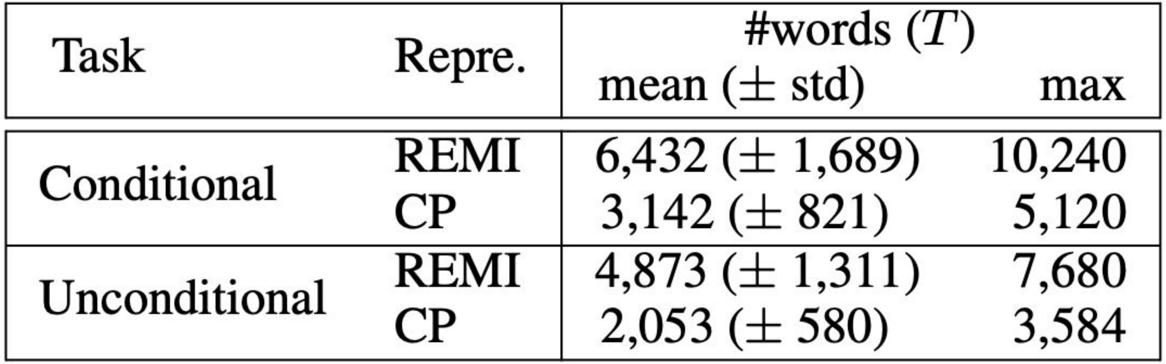
⬜ Given

🟩 Autoregressive generation



Experiments

- **Sequence Length**

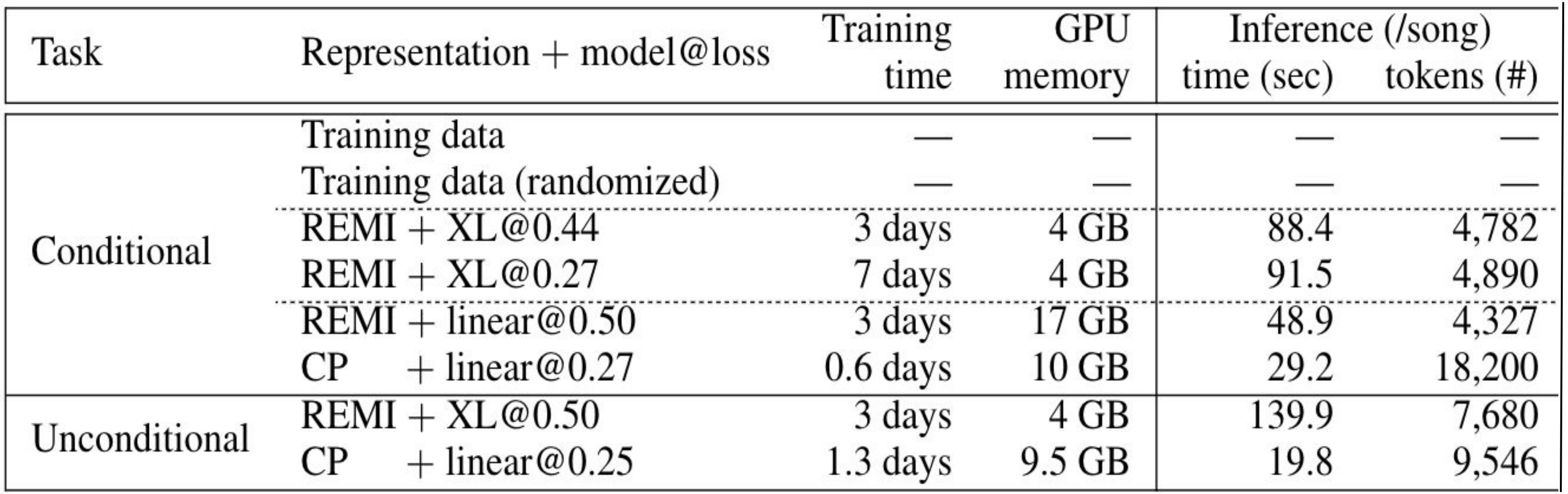


Task	Repre.	#words (T)	
		mean (± std)	max
Conditional	REMI	6,432 (± 1,689)	10,240
	CP	3,142 (± 821)	5,120
Unconditional	REMI	4,873 (± 1,311)	7,680
	CP	2,053 (± 580)	3,584

- **Backbone Models**
  - **Transformer-XL:** recurrence
  - **Linear Transformer:** kernelization

Quantitative Evaluation

- Single GPU (2080ti, 11GB)
- Training & Inference Time



Task	Representation + model@loss	Training time	GPU memory	Inference (/song) time (sec)	Inference (/song) tokens (#)
Conditional	Training data	—	—	—	—
	Training data (randomized)	—	—	—	—
	REMI + XL@0.44	3 days	4 GB	88.4	4,782
	REMI + XL@0.27	7 days	4 GB	91.5	4,890
	REMI + linear@0.50	3 days	17 GB	48.9	4,327
	CP + linear@0.27	0.6 days	10 GB	29.2	18,200
Unconditional	REMI + XL@0.50	3 days	4 GB	139.9	7,680
	CP + linear@0.25	1.3 days	9.5 GB	19.8	9,546

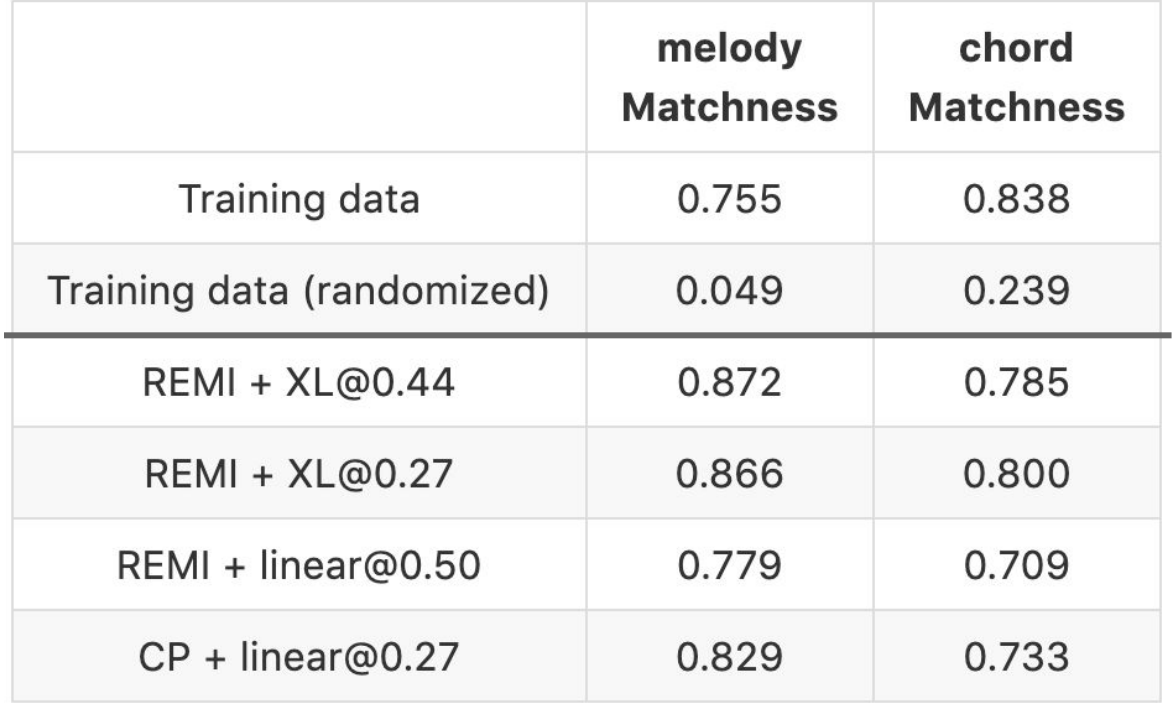
Quantitative Evaluation

- Metrics for conditional generation
    - Melody matchness
        - Bar-wise Longest Common Sub-sequence (LCS) Matchness

$$Matchness_{Melody} = \frac{|LCS(Seq_{Melody}, Seq_{Piano})|}{|Seq_{Melody}|}$$

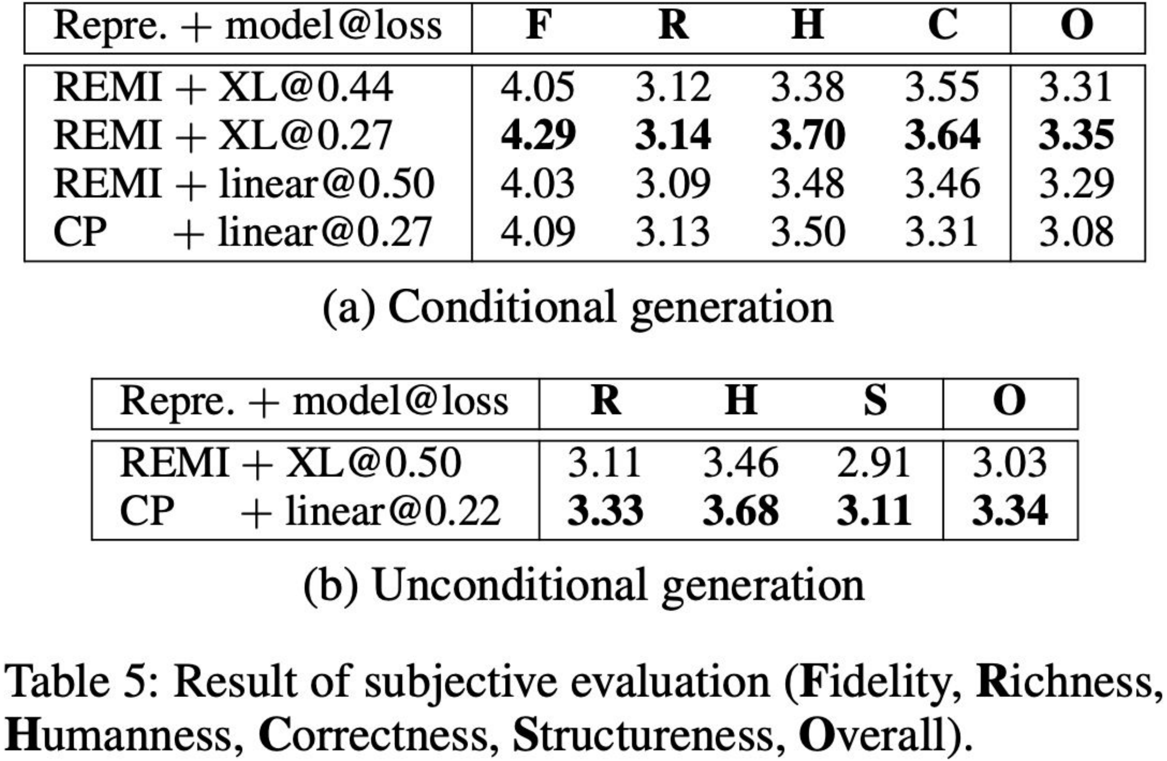
-
    - Chord matchness
        - Segmentwise Cosine Similarity of chormagramss

Quantitative Evaluation



	melody Matchness	chord Matchness
Training data	0.755	0.838
Training data (randomized)	0.049	0.239
REMI + XL@0.44	0.872	0.785
REMI + XL@0.27	0.866	0.800
REMI + linear@0.50	0.779	0.709
CP + linear@0.27	0.829	0.733

User Study



Repre. + model@loss	**F**	**R**	**H**	**C**	**O**
REMI + XL@0.44	4.05	3.12	3.38	3.55	3.31
REMI + XL@0.27	**4.29**	**3.14**	**3.70**	**3.64**	**3.35**
REMI + linear@0.50	4.03	3.09	3.48	3.46	3.29
CP + linear@0.27	4.09	3.13	3.50	3.31	3.08

(a) Conditional generation

Repre. + model@loss	**R**	**H**	**S**	**O**
REMI + XL@0.50	3.11	3.46	2.91	3.03
CP + linear@0.22	**3.33**	**3.68**	**3.11**	**3.34**

(b) Unconditional generation

Table 5: Result of subjective evaluation (**F**idelity, **R**ichness, **H**umanness, **C**orrectness, **S**tructureness, **O**verall).

Summary

- Long sequence modeling
  - memory-efficient transformer
  - compact represetation
- Compound Word Transformer
- Faster in training and inference, with comparable performance

Summary

- Full Song Generation?
    - EOS token generaion
        - transformer-XL ❌
        - linear transformer ⭕
    - Structural Pattern? (like AABA forms)
        - still No

Demo

- Sound Cloud
  - https://soundcloud.com/yating_ai/sets/compound-word-transformer-demo

- Github
  - https://github.com/YatingMusic/compound-word-transformer